Introduction Supervised Learning
IMA205

Pietro Gori

Deadline: Upload the answers and the two notebooks to E-campus. Please verify the deadline on E-Campus.

Theoretical questions

OLS

We have seen that the OLS estimator is equal to $\beta^* = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ which can be rewritten as $\beta^* = H\mathbf{y}$. Let $\tilde{\beta} = C\mathbf{y}$ be another linear unbiased estimator of β where C is a $d \times n$ matrix, e.g. $C = H + D$ where D is a non-zero matrix .

- Demonstrate that OLS is the estimator will the smallest variance: compute $\mathbf{E}[\tilde{\beta}]$ and $\text{Var}(\tilde{\beta}) = \mathbf{E}[(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])^T]$ and show when and why $\text{Var}(\beta^*) < \text{Var}(\tilde{\beta})$. Which assumption of OLS do we need to use ?

Ridge regression

Suppose that both $\mathbf{y}$ and the columns of $\mathbf{x}$ are centered ($\mathbf{y}_c$ and $\mathbf{x}_c$) so that we do not need the intercept β_0. In this case, the matrix $\mathbf{x}_c$ has d (rather than d+1) columns. We can thus write the criterion for ridge regression as:

$$\beta^*_{ridge} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda||\beta||_2^2 \tag{1}$$

- Show that the estimator of ridge regression is biased (that is $\mathbf{E}[\beta^*_{ridge}] \neq \beta$).
- Recall that the SVD decomposition is $\mathbf{x}_c = UDV^T$. Write down by hand the solution β^*_{ridge} using the SVD decomposition. When is it useful using this decomposition ? Hint: do you need to invert a matrix ?
- Remember that $\text{Var}(\beta^*_{OLS}) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$. Show that $\text{Var}(\beta^*_{OLS}) \geq \text{Var}(\beta^*_{ridge})$.
- When λ increases what happens to the bias and to the variance ? Hint: Compute MSE = $\mathbf{E}[(y_0 - x_0^T\beta^*_{ridge})^2]$ at the test point (x_0, y_0) with $y_0 = x_0^T\beta + \epsilon_0$ being the true model and $x_0^T\beta^*_{ridge}$ the ridge estimate.
- Show that $\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1+\lambda}$ when $\mathbf{x}_c^T\mathbf{x}_c = I_d$

OLS

(1) $\beta^* = Hy$ where $H = (X^T X)^{-1} X^T$

β^*: OLS estimator

Let $\tilde{\beta} = Cy$ where C is $d \times n$ matrix $C = H + D$, where D is non-zero

$\tilde{\beta}$: another estimator

- Demonstrate that OLS is the estimator with the smallest Var

- $\mathbb{E}[\beta^*] = H \cdot \mathbb{E}[y]$
- $\text{Var}(\beta^*) = \mathbb{E}[(\beta^* - \mathbb{E}[\beta^*])(\beta^* - \mathbb{E}[\beta^*])^T]$

$$= \mathbb{E}[H(y - \mathbb{E}[y])(H(y - \mathbb{E}[y])^T)]$$

$$= H \,\text{Var}(y)\, H^T$$

- $\mathbb{E}[\tilde{\beta}] = C \cdot \mathbb{E}[y]$
- $\text{Var}(\tilde{\beta}) = \mathbb{E}[(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])^T]$

$$= \mathbb{E}[(\tilde{\beta} - C \cdot E[y])(\tilde{\beta} - C \cdot E[y])^T]$$

$$= \mathbb{E}[(C \cdot y - C \cdot E[y]) \cdot (Cy - C E[y])^T]$$

$$= \mathbb{E}[C \cdot (y - E[y])\ (C(y - E[y]))^T]$$

$$= \mathbb{E}[C \cdot (y - \mathbb{E}[y])\,(y - \mathbb{E}[y])^T \cdot C^T]$$

$$= C \cdot \mathbb{E}[(y - E[y])(y - \mathbb{E}[y])^T] \cdot C^T$$

$$\text{Var}(\tilde{\beta}) = C \cdot \text{Var}(y) \cdot C^T = (H + D)\,\text{Var}(y)\,(H + D)^T$$

$$\text{Var}(\tilde{\beta}) = \underbrace{H \,\text{Var}(y)\, H^T} + D \,\text{Var}(y)\, D^T$$

$$\text{Var}(\tilde{\beta}) = \text{Var}(\beta^*) + \underbrace{D \,\text{Var}(y)\, D^T}_{\text{positive matrix}} > \text{Var}(\beta^*)$$

Ridge Regression

(2)

$$\beta_n^* = \arg\min_\beta (y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda \|\beta\|^2$$

$$\frac{\partial}{\partial \beta}\left[(y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda \|\beta\|_2^2\right] = 0$$

$$-2X_c^T (y_c - X_c\beta_n^*) + 2\lambda\beta^* = 0$$

$$-2X_c^T y_c + 2X_c^T X_c \beta_n^* + 2\lambda\beta^* = 0$$

$$-X_c^T y_c + X_c^T X_c \beta_n^* + \lambda\beta_n^* = 0$$

$\lambda > 0$ always invertible

$$X_c^T y_c = X_c^T X_c \beta_n^* + \lambda\beta_n^* = (X_c^T X_c + \lambda I)\beta_n^*$$

$$\beta_n^* = (X_c^T X_c + \lambda I)^{-1} X_c y_c \;//$$

$$\mathbb{E}[\beta_n^*] = \mathbb{E}[-(X_c^T X_c + \lambda)^{-1} X_c y_c] = (X_c^T X_c + \lambda)^{-1} X_c \mathbb{E}[y_c] \;//$$

$$\mathbb{E}[\beta_n^*] \neq \beta$$

(3) $x_c = UDV^T$

↳ From the previous question

$$\beta_R^* = (X_c^T X_c + \lambda I)^{-1} X_c^T y_c$$

$$\beta_R^* = [(UDV^T)^T (UDV^T) + \lambda I]^{-1} (UDV^T)^T y_c$$

$$\beta_R^* = [V D^T \underbrace{U^T U}_{I} D V^T + \lambda I]^{-1} V D^T U^T y_c$$

Since D is a diagonal matrix $D^T = D$

$$\beta_R^* = [V D^2 V^T + \lambda I]^{-1} V D U^T y_c$$

Since V is an orthonormal matrix $V^T = V^{-1} \Rightarrow \underline{V V^T = I}$ //

$$\beta_R^* = [V D^2 V^T + \lambda I V V^T]^{-1} V D U^T y_c$$

$$\beta_R^* = [V (D^2 + \lambda I) V^T]^{-1} V D U^T y_c$$

$$\beta_R^* = (V^T)^{-1} (D^2 + \lambda I)^{-1} V^{-1} V D U^T y_c$$

$$\boxed{\beta_R^* = V (D^2 + \lambda I)^{-1} D U^T y_c}$$

↳ Since D is a diagonal matrix, $D^2 + \lambda I$ is also a diagonal matrix. It means that the inverse matrix is a diagonal matrix with elements $e_{d-1} = \frac{1}{e_d^2 + \lambda}$, where e_d are the elements from the D matrix. So, it's not more necessary calculate the inverse matrix wich could be a computational costly operation.

(4) $\mathrm{Var}(\beta^*_{OLS}) = \sigma^2 (X^T X)^{-1}$,

Show that $\mathrm{Var}(\beta^*_n) \leq \mathrm{Var}(\beta^*_{OLS})$

$$\beta^*_n = (X_c^T X_c + \lambda I)^{-1} X_c^T y_c$$

$$\mathbb{E}[\beta^*_n] = (X_c^T X_c + \lambda I)^{-1} X_c^T \mathbb{E}[y_c]$$

$$\mathrm{Var}(\beta^*_n) = \mathbb{E}\left[(\beta^*_n - \mathbb{E}[\beta^*_n])(\beta^*_n - \mathbb{E}[\beta^*_n])^T\right]$$

→ By means of notation let's assume $H = (X_c^T X_c + \lambda I)^{-1} X_c^T$

So,

$$\mathrm{Var}(\beta^*_n) = \mathbb{E}\left[(Hy_c - H\mathbb{E}[y_c])(Hy_c - H\mathbb{E}[y_c])^T\right]$$

$$= \mathbb{E}\left[H(y_c - \mathbb{E}[y_c])(y_c - E[y_c])^T H^T\right]$$

$$= H \cdot \underbrace{\mathbb{E}\left[(y_c - \mathbb{E}[y_c])(y_c - \mathbb{E}[y_c])^T\right]}_{\mathrm{Var}(y) = \sigma^2} H^T$$

$$\Rightarrow \mathrm{Var}(\beta^*_n) = \sigma^2 HH^T = \sigma^2 \left[(X_c^T X_c + \lambda I)^{-1} X_c^T\right]\left[(X_c^T X_c + \lambda I)^{-1} X_c^T\right]^T$$

$$\mathrm{Var}(\beta^*_n) = \sigma^2 (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

Since $\lambda \geq 0$ and $X_c^T X_c \geq 0$, we can say that

$$(X_c^T X_c)^{-1}\left[(X_c^T X_c)^{-1}\right]^T \geq (X_c^T X_c + \lambda I)^{-1}\left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

So, is also true:

$$(X_c^T X_c)^{-1} X_c^T X_c \left[(X_c^T X_c)^{-1}\right]^T \geq (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

$$\left[(X_c^T X_c)^{-1}\right]^T \geq (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

Also true for the transpose matrix

$$(X_c^T X_c)^{-1} \geq (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

$$\sigma^2 (X_c^T X_c)^{-1} \geq \sigma^2 (X_c^T X_c + \lambda I)^{-1} X_c^T X_c \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

Finally

$$\text{Var}(\beta^*_{OLS}) \geq \text{Var}(\beta^*_R) \;//$$

(5) $\lambda \to \infty$ $Var(\beta_R^*) \to ?$ $Bias(\beta_R^*) \to ?$

$$Var(\beta_R^*) = \underline{\sigma^2}(X_c^T X_c + \lambda I)^{-1} \underline{X_c^T X_c} \left[(X_c^T X_c + \lambda I)^{-1}\right]^T$$

↳ independent from lambda

When $\lambda \to \infty$ $Var(\beta_R^*) \to 0$

For $\lambda \to \infty$ $\beta_R^* \to 0$,

$$MSE = \mathbb{E}\left[(y_0 - x_0^T \beta_R^*)^2\right]$$

$$y_0 = x_0^T \beta + \varepsilon$$

$$MSE = \mathbb{E}\left[(x_0^T(\beta - \beta_R^*) + \varepsilon_0)^2\right]$$

$$MSE = \mathbb{E}\left[(x_0^T(\beta - \beta_R^*))^2 + 2(x_0^T(\beta - \beta_R^*))\varepsilon_0 + \varepsilon_0^2\right]$$

↳ Let's assume $\varepsilon_0 \sim N(0, Var(\varepsilon_0))$

$$MSE = (x_0^T(\beta - \beta_R))^2 + 2(x_0^T(\beta - \beta_R^*))\underbrace{\mathbb{E}[\varepsilon_0]}_{\to 0} + Var(\varepsilon_0)$$

$$MSE = \underbrace{(x_0^T(\beta - \beta_R))^2}_{bias^2} + Var(\varepsilon_0)$$

With $\lambda \to \infty$, $Var \to 0$, $\beta_R \to 0$ wich means that the bias term increases.

⑥ Show that $\beta_n^* = \frac{\beta_{OLS}}{1+\lambda}$ when $X_c^T X_c = I$

$\beta_n^* = (X_c^T X_c + \lambda I)^{-1} X_c^T y_c$ $\qquad$ $\beta_{OLS} = (X_c^T X_c)^{-1} X_c^T y_c$

$X_c^T X_c = I \Rightarrow \beta_{OLS}^* = X_c^T y_c$

$\Rightarrow \beta_n^* = (I + \lambda I)^{-1} X_c^T y_c$

$$\boxed{\beta_n^* = \frac{\beta_{OLS}^*}{1+\lambda}}$$

//

Elastic Net

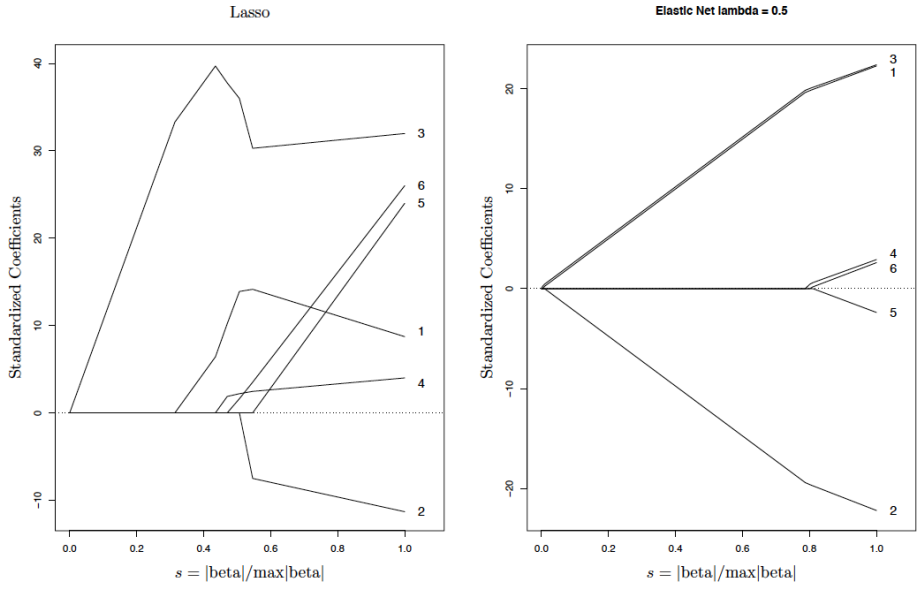
Using the previous notation, we can also combine Ridge and Lasso in the so-called Elastic Net regularization :

$$\beta^*_{ElNet} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1 \tag{2}$$

Calling $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, solving the previous Eq. is equivalent to:

$$\beta^*_{ElNet} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda\left[\alpha(\sum_{j=1}^{d}\beta_j^2) + (1-\alpha)(\sum_{j=1}^{d}|\beta_j|)\right] \tag{3}$$

- This regularization overcomes some of the limitations of the Lasso, notably:
  - If $d > N$ Lasso can select at most N variables → ElNet removes this limitation
  - If a group of variables are highly correlated, Lasso randomly selects only one variable → with ElNet correlated variables have a similar value (grouped)
  - Lasso solution paths tend to vary quite drastically → ElNet regularizes the paths
  - If $N > d$ and there is high correlation between the variables, Ridge tends to have a better performance in prediction → ElNet combines Ridge and Lasso to have better (or similar) prediction accuracy with less (or more grouped) variables



- Compute by hand the solution of Eq.2 supposing that $\mathbf{x}_c^T\mathbf{x}_c = I_d$ and show that the solution is: $\beta^*_{ElNet} = \frac{(\beta^*_{OLS})_j \pm \frac{\lambda_1}{2}}{1+\lambda_2}$

Elastic Net

⑦ $X_c^T X_c = I$ $\quad \beta_E^* = \dfrac{\beta_{OLS}^* \pm \lambda_1/2}{1+\lambda_2}$ $\quad \beta_{OLS} = (X_c^T X_c)^{-1} X_c^T y_c = X_c^T y_c$

$$\beta_E^* = \arg\min_\beta (y_c - X_c\beta)^T (y_c - X_c\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$$

$$\frac{\partial}{\partial \beta_E^*} = -2X_c^T(y_c - X_c\beta_E^*) + 2\lambda_2\beta_E^* + \lambda_1 \begin{cases} -1, & \beta<0 \\ [-1,1], & \beta=0 \\ 1, & \beta>1 \end{cases} = 0$$

subgradient

21.1

$$-2\underbrace{X_c^T y_c}_{\beta_{OLS}^*} - 2\underbrace{X_c^T X_c}_{I}\beta_E^* + 2\lambda_2\beta_E^* \pm \lambda_1 = 0$$

$$-2\beta_{OLS}^* - 2\beta_E^*(1-\lambda_2) \pm \lambda_1 = 0$$

$$\beta_{OLS}^* + \beta_E^*(1-\lambda_2) \mp \frac{\lambda_1}{2} = 0$$

$$\beta_E^*(1-\lambda_2) = -\beta_{OLS}^* \pm \frac{\lambda_1}{2} = 0$$

$$\beta_E^* = \frac{\beta_{OLS}^* \pm \frac{\lambda_1}{2}}{\lambda_2 - 1} \quad //$$