Tutorial & Practical 4: Model Selection

Question 1

In this question we are interested in deriving an algorithm for solving Lasso.

Given the model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2 I_n)$. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained by least square estimation.

1. Let $y \in \mathbb{R}$, find the solution $u \in \mathbb{R}$ that minimizes

$$(y-u)^2 + \lambda|u|$$

2. Plot the solution as a function of y

3. Use this solution to derive an algorithm for solving

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|$$

Question 2

Let $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)^\top$ be an $n \times q$ matrix of observations for which we postulate the parametric model

$$Y = XB + \boldsymbol{\varepsilon} \quad \text{where} \quad \text{vec}(\boldsymbol{\varepsilon}) \sim \mathcal{N}\left(0, \Sigma \otimes I_n\right)$$

where X is a known $n \times k$ design matrix of rank k, B is a $k \times q$ matrix of unknown parameters, $\boldsymbol{\varepsilon}$ is the $n \times q$ matrix of errors and Σ is a $q \times q$ matrix of error covariance.

1. Give the expression of the log likelihood

2. Find the number of parameters involved in the model

3. Derive the expressions of AIC and BIC

Question 3

Let $\mathbf{y} \in \mathbb{R}^n$ be a vector of observation for which we postulate the linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n \sim \mathcal{N}(0, \sigma^2 I_n)$. The dimension k of $\boldsymbol{\beta}$ is estimated using AIC.

1. Give the form of the AIC criterion

2. Derive the expression of the probability of overfitting.