Laboratorium 5 : Mean squared error, mean squared prediction error

Patrick Tardivel, Wrocław university

Notations:

- The notation $\mathcal{N}(\mu, \Sigma)$ represents the distribution of a Gaussian vector whose expected value is μ and whose covariance matrix is Σ.
- The notation Id_q represents the $q \times q$ identity matrix.

Exercise 1 *Let Z be a Gaussian vector having a $\mathcal{N}(\mu, Id_n)$ distribution. What is the mean squared error of Z that is what is the value of $E(\|Z-\mu\|_2^2)$? The James-Stein estimator of μ is defined by $\widehat{\mu}_{JS} = Z-(n-2)Z/\|Z\|_2^2$. When $n = 10$ and $\mu_t = (t, \dots, t) \in \mathbb{R}^n$, by simulating lot of times Z, draw the curve of the function $t > 0 \mapsto \mathbb{E}(\|\widehat{\mu}_{JS} - \mu_t\|_2^2)$. Add to this figure the horizontal line $y = E(\|Z - \mu\|_2^2)$. What do you notice?*

In the following exercise, we consider the linear Gaussian regression model $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 Id_n)$. We remind the following facts:

- The LASSO estimator is defined as follows

$$\widehat{\beta}(\lambda) := \operatorname*{argmin}_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_1.$$

- Given $\lambda > 0$, the random variable $\|Y - X\widehat{\beta}(\lambda)\|_2^2 + 2\sigma^2\|\widehat{\beta}(\lambda)\|_0 - n\sigma^2$ is an unbiased estimator of the LASSO mean square prediction error ($\mathbb{E}(\|X\widehat{\beta}(\lambda) - X\beta\|_2^2)$ given by the SURE formula.
- When X is orthogonal, namely when $X'X = Id_p$, the LASSO estimator has the following expression

$$\forall i \in \{1, \dots, p\}, \widehat{\beta}_i(\lambda) := \operatorname{sign}(\widehat{\beta}_i^{\text{ols}})(|\widehat{\beta}_i^{\text{ols}}| - \lambda)_+, \text{ where } \widehat{\beta}^{\text{ols}} = X'Y \text{ and } (t)_+ = \max\{t, 0\}.$$

A natural way to select the tuning parameter λ in order to have a small mean squared prediction error for the LASSO is to minimize the function

$$\lambda > 0 \mapsto \|Y-X\widehat{\beta}(\lambda)\|_2^2+2\sigma^2\|\widehat{\beta}(\lambda)\|_0-n\sigma^2, \text{ (or equivalently to minimize } \lambda > 0 \mapsto \|Y-X\widehat{\beta}(\lambda)\|_2^2+2\sigma^2\|\widehat{\beta}(\lambda)\|_0).$$

The following exercise compare LASSO estimator when 1) the tuning parameter is selected in order to minimize the mean squared prediction error with 2) when the tuning parameter is selected in order to control the familywise error rate.

Exercise 2 *Let $X \in \mathbb{R}^{n\times p}$ be an orthogonal matrix and λ_0 be the $(1 + \sqrt[p]{1-\alpha})/2$ quantile of a $\mathcal{N}(0, \sigma^2)$ distribution.*

1) *Prove the following inequality*

$$\mathbb{P}\left(\exists i \notin \operatorname{supp}(\beta) \text{ such that } \widehat{\beta}_i(\lambda_0) \neq 0\right) \leq \alpha.$$

This probability (called the familywise error rate) is the probability to do not correctly estimate at 0 with LASSO estimator at least one null component of β. In which case this inequality is an equality?

Now, let us set $n = 10, p = 5, \sigma = 1, \alpha = 0.05, \beta = (3, 1, 0, 0, 0)$ *and let us set* X *and* ε *as follows :*

library(pracma)
set.seed(2020)
$X = randortho(10)[, 1:5]$
$\varepsilon = rnorm(10)$

2) *Given a particular observation* Y *(associated to the particular observation of* ε*), draw the function*

$$\lambda > 0 \mapsto \|Y - X\widehat{\beta}(\lambda)\|_2^2 + 2\|\widehat{\beta}(\lambda)\|_0 - 5.$$

Add on your figure vertical lines $x = |\widehat{\beta}_i^{\text{ols}}|$ *for* $i \in \{1, \ldots, 5\}$ *and determine graphically the value* λ_1 *for which the estimation of the mean squared prediction error for the LASSO is minimal.*

3) *By simulating lot of observations of* ε *(thus lot of observations of* Y*) compare* $\mathbb{E}(\|X\widehat{\beta}(\lambda_1) - X\beta\|^2)$ *with* $\mathbb{E}(\|X\widehat{\beta}(\lambda_0) - X\beta\|^2)$*. For which tuning parameter* $\lambda \in \{\lambda_0, \lambda_1\}$ *the mean square error is minimal?*

4) *By simulating lot of observations of* ε*, compare probabilities* $\mathbb{P}(\exists i \notin \text{supp}(\beta) \text{ such that } \widehat{\beta}_i(\lambda_1) \neq 0)$ *with* $\mathbb{P}(\exists i \notin \text{supp}(\beta) \text{ such that } \widehat{\beta}_i(\lambda_0) \neq 0)$*. For which tuning parameter* $\lambda \in \{\lambda_0, \lambda_1\}$ *this probability (the familywise error rate) is the smallest one?*