Année 2014/2015 UNIVERSITÉ TOULOUSE I CAPITOLE

Big Data

TP 2 **High dimensional regression**

We remind shorty the formalism of penalized linear regression. We consider the following model

$$Y = X\theta_0 + \epsilon.$$

The noise is assumed to be centered and we aim to estimate θ_0.

In this exercice, we will consider three penalized methods: Ridge regression, Elastic Net, and the Lasso. The general form of these estimators solve the following optimization problem

$$\hat{\theta}_\lambda := \arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|_2^2 + pen(\theta, \lambda). \tag{1}$$

We define the penalties as

- $pen(\theta, \lambda) = \lambda \|\theta\|_2^2$ (Ridge regression / Tikhonov regularization)
- $pen(\theta, \lambda) = \lambda \|\theta\|_1$ (Lasso)
- $pen(\theta, \lambda) = \lambda \|\theta\|_1 + \mu \|\theta\|_2^2$ (Elastic Net).

The penalization parameter λ is a positive number for the Ridge and Lasso regression. It is a 2-dimensional vector of $\mathbb{R}_+^2$ in the case of the Elastic Net.

Exercice 1. [Function for the generation of the data]

For the next experiments, it will be important to code a function that generates some data $(X_i, Y_i)_{1 \leq i \leq n}$. This R/matlab function should take in argument n, p as well as an integer L, K and a real vector θ.

1. We define a squared covariance matrix Σ of size $p \times p$, the structure is block-diagonal with K blocks of size $\ell \times \ell$. Draw the matrix Σ.
2. Each block is referred to as Σ_{jj} and is defined as

$$\Sigma_{jj} = \frac{j-1}{K} \mathbf{1}_{\ell \times \ell} + \frac{K+1-j}{K} Id_\ell.$$

   Write a R function or Matlab function that generates a matrix Σ_{jj}, as well as a whole matrix Σ.

3. Conclude with a R function or Matlab function that generates n independent replications of (X, Y) with

$$X \sim \mathcal{N}(0, \Sigma)$$

   and

$$Y = X\theta_0 + \epsilon, \qquad \text{where} \qquad \epsilon \perp X.$$

Exercice 2. [Generation of the data]
For each situation, build the data with using the previous code and identify if the data are sparse or dense, with correlated entries or not.

1. $n = 100, K = 1, \ell = 50$ and
$$\theta_0 = (\underbrace{-2, \ldots - 2}_{\ell\,\text{times}}, \underbrace{+2 \ldots, +2}_{(K-1)\times\ell\,\text{times}})$$

2. $n = 100, K = 1, \ell = 50$ and
$$\theta_0 = (\underbrace{-2, \ldots, -2}_{5\,\text{times}}, \underbrace{0, \ldots, 0}_{40\,\text{ times}}, \underbrace{2, \ldots 2}_{5\,\text{times}}).$$

3. $n = 100, K = 3, \ell = 50$ and
$$\theta_0 = (\underbrace{-2, \ldots - 2}_{\ell\,\text{times}}, \underbrace{+3 \ldots, +3}_{(K-1)\times\ell\,\text{times}})$$

4. $n = 100, K = 3, \ell = 50$ and
$$\theta_0 = (\underbrace{-2, \ldots, -2}_{5\,\text{times}}, \underbrace{0, \ldots, 0}_{40\,\text{ times}}, \underbrace{2, \ldots 2}_{5\,\text{times}}).$$

Exercice 3. [Solving the Ridge regression]

1. In R: RIDGE function of the package MASS (use the lm.ridge function).
2. In Matlab: Ridge function has been previously built in the first session.

- Compute the ridge regression $\hat{\theta}(\lambda)$ for 30 different values of the regularization parameter λ according to a geometric grid:
$$\lambda_j = e^j$$
- Plot on the same graph the evolution of the several values of the estimator $\lambda \mapsto (\theta_j(\lambda))_{1\leq j\leq p}$
- The reading `http://stat.genopole.cnrs.fr/_media/members/jchiquet/teachings/ridge.pdf` may be helpful.

Exercice 4. [Solving the Lasso regression]

1. In R: Use the LARS package and especially the function LARS. The useful objects are *object_lasso$lambda* where *object_lasso*=lars(X,Y,type="lasso").
2. In Matlab: Lasso is already developed here: `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897` or in the statistical toolbox. Download either this package or use the statistical toolbox.

- Compute the Lasso regression $\hat{\theta}(\lambda)$ for 30 different values of the regularization parameter λ according to a geometric grid:
$$\lambda_j = e^j$$
- Plot on the same graph the evolution of the several values of the estimator $\lambda \mapsto (\theta_j(\lambda))_{1\leq j\leq p}$

Exercice 5. **[Solving the Elastic Net regression]**

1. In R: Use the ENET package and especially the function ENET. It is also possible to use the GLMNET package.
2. In Matlab: Elastic Net is already developed here: `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897` or in the statistical toolbox. Download either this package or use the statistical toolbox.

Exercice 6. **[Evaluation]**

- In each of the 4 situations, rank the three regression methods.
- We may be interested in the prediction error, as well as the support recovery abilities or the estimation errors. Recall the definitions of these three criteria.
- Use a cross-validation method or a test set to estimate these errors.