# Sparsité, Estimation et Sélection de Variables
# LASSO, Ridge et Elastic Net

*Evgenii Chzhen & Henry VONG*

*24 november 2015.*

## Contents



### Introduction

Consider the following linear model :

$$Y = X\beta + \eta,$$

where $\eta \overset{\text{i.i.d}}{\sim} \mathbb{N}(0, I_n)$, $\beta \in \mathbb{R}^p$ and $X_{.,i}$ are i.i.d. standart gaussian vectors for each $i \in 1, ..., p$. We study three estimators which are defined as it follows :

1. **LASSO**
$$\hat{\beta}^L = \underset{\beta}{\text{argmin}} \Big[\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\Big],$$
for $\lambda > 0$
2. **Ridge**
$$\hat{\beta}^R = \underset{\beta}{\text{argmin}} \Big[\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2 + \frac{\mu}{2}\|\beta\|_2\Big],$$
for $\mu > 0$
3. **Elastic net**
$$\hat{\beta}^{EN} = \underset{\beta}{\text{argmin}} \Big[\frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2 + \lambda\Big(\frac{1-\alpha}{2}\|\beta\|_2 + \alpha\sum_{j=1}^{p}|\beta_j|\Big)\Big],$$
for $\lambda > 0$ and $\alpha \in [0, 1]$. One can notice that $\alpha = 0$ is equivalent to Ridge and $\alpha = 1$ is equivalent to LASSO.

For more details and theoretical background behind theese methodes see ((Tsybakov 2008, 59), (T. Zou H. & Hastie 2005, 301–7), (H. Zou H. 2004)).

## 1. First dataset

Consider the following linear model :

$$Y = X\beta + \eta,$$

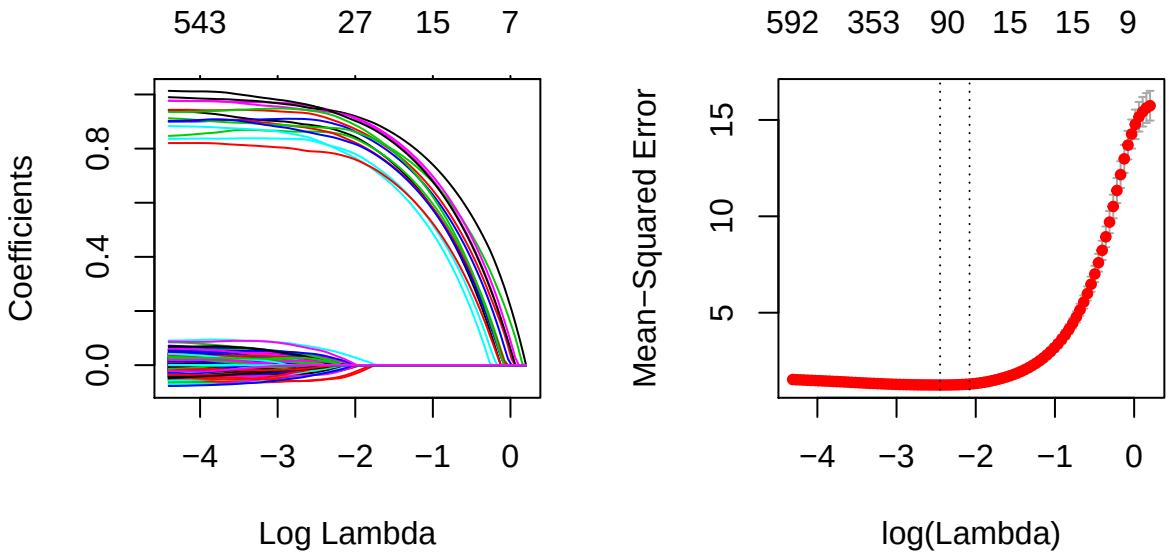
where $\eta \overset{\text{i.i.d}}{\sim} \mathbb{N}(0, I_n)$, $\beta \in \mathbb{R}^p$ and $X_{\cdot,i}$ are i.i.d. standart gaussian vectors for each $i \in 1, ..., p$. We put $n = 1000$, $p = 5000$,

$$\beta_1 = ... = \beta_{15} = 1,$$

$$\beta_{16} = ... = \beta_{5000} = 0.$$

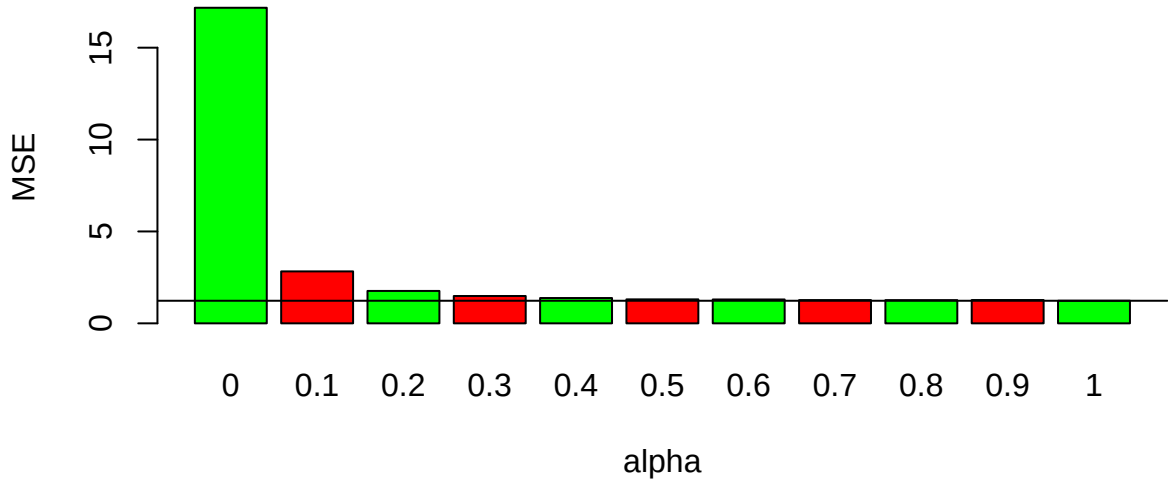
For each value of $\alpha \in \{0, 0.1, 0.2, ..., 0.9, 1\}$ we compute $\lambda_{1se}$ by cross validation, where $\lambda_{1se}$ largest value of $\lambda$ such that error is within 1 standard error of the minimum. To choose an optimal $\alpha$ we separate our dataset on 2 groups namely train set and test set, after fitting a single model on train set we compute MSE on test set and choose $\alpha$ which corresponds to the minimal value of error.

Coefficients path and a plot of cross-validated MSE in case of LASSO




**Corollary**: One can see on left plot that LASSO is able to zero out a coefficients. As Lambda increases, MSE increases rapidly. The coefficients are reduced too much and they do not adequately fit the responses.

Obtained errors on testing set :




One can see that the best is alpha = 1, which is equivalent to LASSO.

## 2. Second dataset

Consider the following linear model :

$$Y = X\beta + \eta,$$

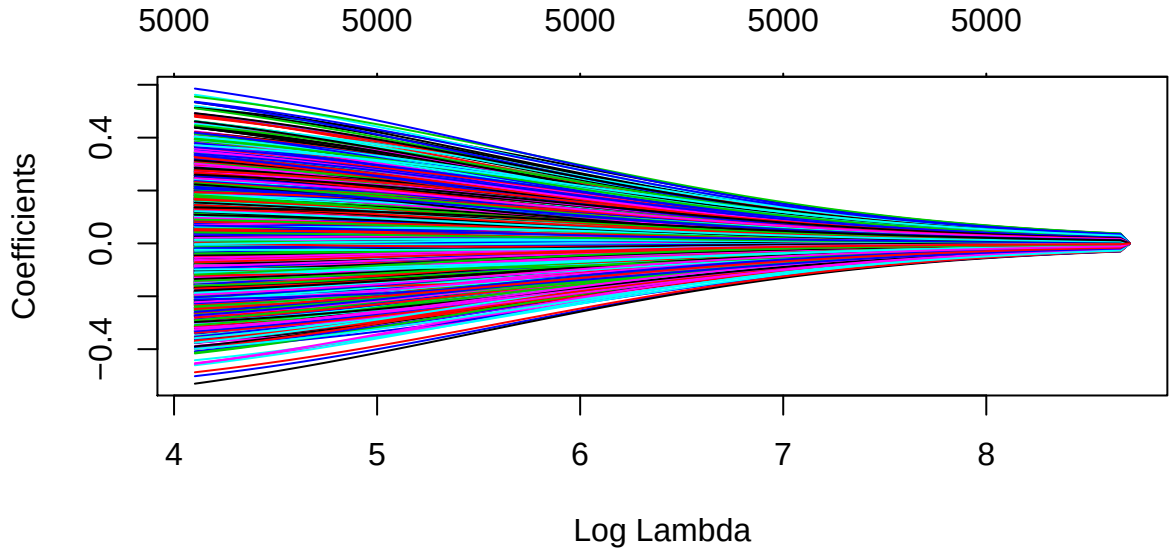
where $\eta \overset{\text{i.i.d}}{\sim} \mathbb{N}(0, I_n)$, $\beta \in \mathbb{R}^p$ and $X_{\cdot,i}$ are i.i.d. standart gaussian vectors for each $i \in 1, ..., p$. We put $n = 1000$, $p = 5000$,

$$\beta_1 = ... = \beta_{1500} = 1,$$

$$\beta_{1501} = ... = \beta_{5000} = 0.$$

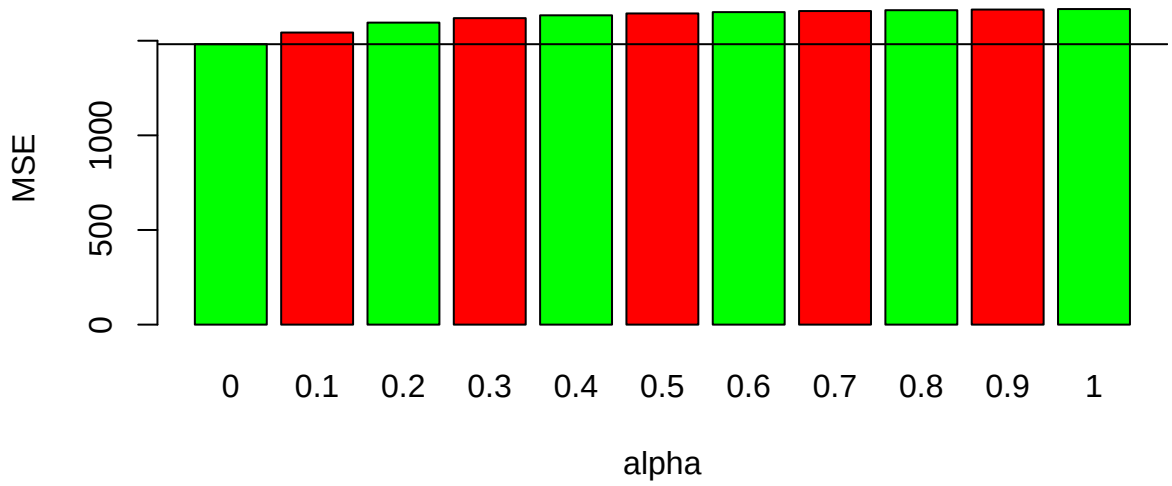
We proceed the same procedure to choose best value of $\alpha$.

Coefficients path in case of RIDGE




**Corollary**: One can see on left plot that Ridge can't shrink coefficients so you include in final model all the coefficients or none of them.

Errors on testing set :




One can see that the best is alpha = 0, which is equivalent to Ridge.

## 3. Third dataset

Consider the following linear model :

$$Y = X\beta + \eta,$$

where $\eta \overset{\text{i.i.d}}{\sim} \mathbb{N}(0, I_n)$, $\beta \in \mathbb{R}^p$ and $X_{.,i} \overset{\text{i.i.d}}{\sim} \mathbb{N}\Big(0, \Sigma\Big)$ vectors for each $i \in 1, ..., p$, where $\Sigma_{pk} = 0.7$ if $k \neq p$ and $\Sigma_{ii} = 1$. We put $n = 100$, $p = 50$,

$$\beta_1 = \beta_2 = 10,$$

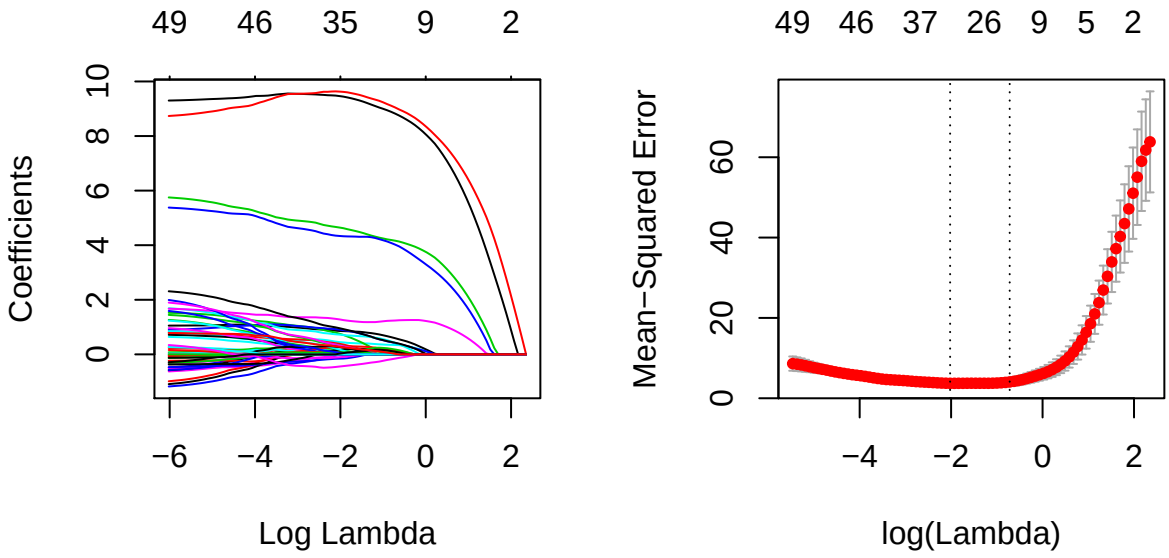
$$\beta_3 = \beta_4 = 5,$$

$$\beta_5 = ... = \beta_{14} = 1,$$

$$\beta_{15} = ... = \beta_{50} = 0.$$

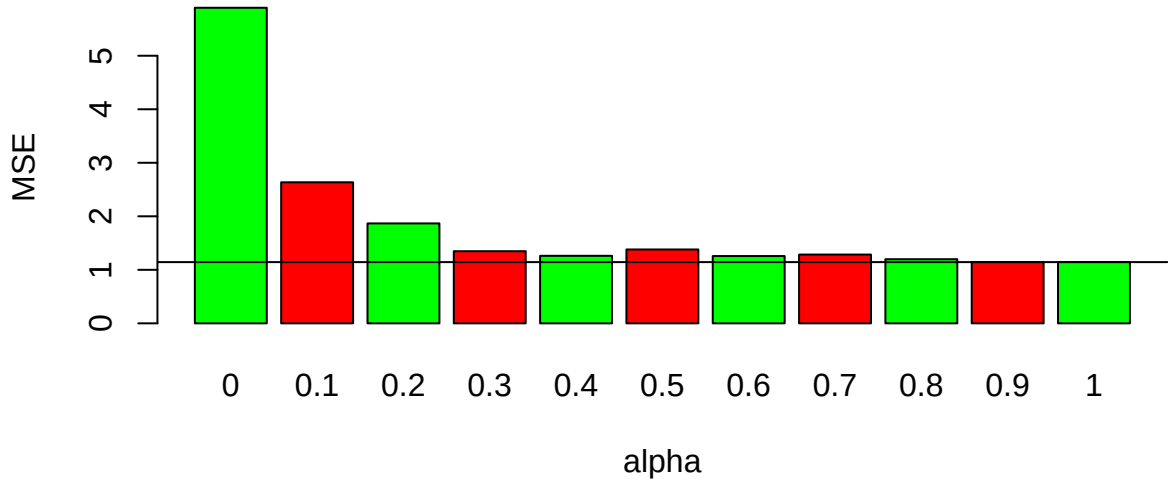
We proceed the same procedure to choose best value of $\alpha$.

Coefficients path and a plot of cross-validated MSE in case of Elastic Net (0.5)




**Corollary**: We can notice almost the same behaviour as in case of LASSO.

Errors on testing set :




one can notice that the best alpha = 0.9, which is equivalent to Elastic net.

## Corollary

One can notice that LASSO is working good in a very sparse situations, whenever RIDGE regression is working in less sparse cases, since Ridge can't zero out coefficients; thus, one either end up including all the coefficients in the model, or none of them (however the ridge regression will penalize our coefficients, such that those who are the least efficient in our estimation will "shrink" the fastest.). The model of Elastic Net is appropriate when the variables are highly correlated. Theese results are in high agreement with a previous expiriments and theoretical results, see for instance (T. Zou H. & Hastie 2005, 301–7).

## References

Tsybakov, Alexandre B. 2008. *Introduction to Nonparametric Estimation.* Springer.

Zou, Hastie, H. 2004. *Sparse Principal Component Analysis.* Journal of Computational; Graphical Statistics.

Zou, T., H. & Hastie. 2005. *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society (Series B).