Ridge, Lasso and OLS comparison with simulated data in Python

Dario Alfredo de Falco
Filippo Palandri

November 2023

Abstract

Our analysis consists in applying different linear regression models with simulated data with Python in order to compare them and evaluate their functioning. Giving that we are directly generating the data we are able to evaluate the performance of these regularization methods given different degree of regularization in order to pick the best. Moreover we will be using a resampling method in order to have a better understanding of how the model behaves with higher variability of the data

1 Methods

As we said in the introduction we are willing to compare different linear regression methods for some given degree of regularization. In particular we are performing 3 different models; the standard OLS, the Ridge regression and the Lasso regression. Data are simulated multiple times, the parameters remain the same but the noise changes. Furthermore, we generated the data using a normal distribution and also simulated standard deviations for the errors and the regressors using a chi squared distribution. Finally, the errors are generated in their random component from a normal distribution and are linearly related to the regressors themselves so as to add some heteroskedasticity. This will be properly assesed by the code by computing heteroskedasticity robust standard errors. Given this framework, ridge and Lasso should be able to detect and exclude variables and have nothing to do with the DGP, and given that we are directly simulating the DGP we can exactly determine the goodness of the result. Every method is trained and fit using a k-cross validation with k splits and the average MSE across the splits is computed. Moreover both lasso and ridge are performed with different values of the regularization degree α and the best model is selected and its train and test errors are plotted. Both Ridge and Lasso are shrinkage methods that aim at shrinking toward zero the coefficients of some parameters in order to trade a slightly higher bias w.r.t. classical OLS with a lower variance and thus overall a lower MSE.

1.0.1 Ridge

The ridge coefficients solve the following minimization problem:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1}$$

Moreover differently from Lasso, Ridge has a closed form solution which is:

$$\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y \tag{2}$$

Where n is the number of observations, p is the number of regressors and λ is is a hyperparameter that controls the amount of shrinkage, it can be seen as a penalization to the minimization problem or as a Lagrange multiplier. The higher is λ the higher is the penalization and hence the shrinkage. λ is what we call α in our code.
Note from (1) and (2) that as $\lambda \to 0$, $\hat{\beta}^{\text{ridge}} \to \hat{\beta}^{\text{OLS}}$, while as $\lambda \to \infty$ all the coefficients will be shrinked toward 0.

1.0.2 Lasso

Lasso coefficients differently solves the following problem:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

Lasso minimizes the absolute values and not the squares, in practice this difference allows the Lasso to shrink much more efficiently the coefficients toward zero. All we said for the Ridge can be said for the Lasso, but not that is has a closed form solution.

2 Results

The true DGP is linear in the form:

$$y = 5 + X\beta + \epsilon \tag{4}$$

Then we simulate some random variables that aren't related to the DGP and we concatenate the data in order to get the whole data set on which we estimate the models. Performances are evaluated according to how precise are the estimates w.r.t. the DGP on different test data-sets. For OLS we plot average test-mse and train-mse for only a sample of generated data, the average is computed over the 5 test-sets. For ridge and lasso we plot the average MSE and we get the optimal α. Then we evaluate how sensible is the cross validation with respect to a different sample by simulating multiple times different data. On these different datasets we repeat our analysis in order to extract the best alpha for each dataset and its correspondent test MSE. Our findings are that the best alpha is consistent across simulations and a similar observation holds for the test MSE which is also quite stable. Due to computing limitations and the limited scope of the project we limited the grid of parameters tested and the number of different datasets. Nonetheless, our method seems to be robust in choosing the right alpha and correctly performing model selection (in lasso case) and proper shrinking (Ridge). Actually, these two model end up performing better than the OLS and give nearly indistinguishable prediction errors. The code allows for a direct generalization and extension of the procedure to more sophisticated approaches.

3 Conclusions

In this empirical exercise we tried to fit various model on different data-sets generated by the same DGP. We found that in a context with some variables that are not included in the DGP, lasso and ridge tend to correctly shrink them to zero and outperform the OLS. Moreover, when trying to check for the robustness of the procedure selecting the alpha, we found that almost every time the same parameter is chosen, suggesting that indeed we found the best hyper-parameter. An extension of this exercise would be to run it again with a more extensive grid of parameters and for a larger amount of datasets and different varieties of them. The code presented would easily allow for such

an implementation, and the reason such a direction is not pursued is due to computational constraints and the limited scope of the project itself.