An empirical comparison between PLS, LASSO, Elasticnet and other models for highly correlated data

Genane Youness, Wassim Nassar

An empirical comparison between PLS, Lasso, Elasticnet and other models for highly correlated data.

Genane Youness [a*] and Wassim Nassar [a]

[a] ISSAE- CNAM, Beirut, Lebanon

Keywords: Multicollinearity, Ridge regression, PLS, Lasso, and Elasticnet.

Introduction

We compare seven regression techniques to solve the question of muticollinearity in multiple regression: PCR, PLS regression, with Ridge, Lasso, Lars, Adaptive Lasso and Elasticnet. An application on real data providing from the Lebanese national center for scientific research CNRSL(Centre National de la Recherche Scientifique Liban) on the area of fire in Lebanon forests in 2005 will be made. The comparison of these techniques is discussed along with some of their advantages and disadvantages.

1 Methods of multicollinearity

when we have correlation between predictor variables, some methods have to be used for dealing with this problem: dimension reduction methods such as PCR and PLS regression[1], or shrinkage penalized methods such as Ridge, Lasso proposed by Tibshirani [2], LARS, Adaptive Lasso and Elasticnet, proposed by Zou and Hastie [3] combing the penalty terms of Lasso and Ridge. The goal of our paper is to compare these methods.

2 Empirical data

We apply the seven techniques to 160 burned areas with five factors affecting the behavior of forest fires as elevation, perimeter, mean slope gradient, mean vegetation density (NDVI), and mean evaporation. they were calibrated using the map burned areas extracted from the visual interpretation of satellite images. The variance inflation of mean slope and mean evaporation are very high respectively 13.56 and 20.89.
To validate our models we divide the data into two sets randomly selected: a training sample, to estimate the model, and a test sample to test the good behavior of the model by calculating R^2, the coefficient of determination and the mean square error $\widehat{\sigma_\varepsilon^2}$.

2.1 Empirical Study

In order to compare the different methods, we kept all predictors, they are summarized in the following graph and table:

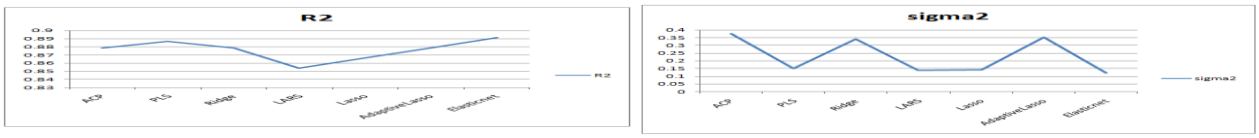

Figure 1. Repartition of the coefficient of determination and the mean square error for all methods

From these results we can choose as best method the Elasticnet having the highest coefficient of determination R^2 (0.891) and the minimum of the means square error (0.121). Note that PLS regression also gives the second best R^2 (0.88).

---

* Corresponding author. E-mail: genane.youness@cnam.fr, Beirut, Lebanon, BP 11- 4661 Beirut, Lebanon.

Table 1. Summary table of estimated parameters of the different methods.

	PCR	PLS	Ridge	LARS	Lasso	Adaptive Lasso	Elasticnet
PERIMETER	0.88	0.88098	0.86825	0.8809797	0.88097972	0.881156354	0.89253770
MEAN_ELEV	0.33	0.334051	0.27028	0.3340512	0.33405121	0.332155233	0.22963326
MEAN_SLOPE	0.048	0.047948	0.08716	0.0479485	0.04794850	0.047307248	0.04549356
MEAN_NDVI	-0.007	-0.00703	-0.03329	-0.0070262	-0.00702625	-0.00599746	-0.04415432
MEAN_EVAPO	0.23	0.228305	0.20763	0.2283054	0.2283053	0.226201495	0.12130083
$\widehat{\sigma_{\varepsilon}^{2}}$	0.375053	0.150727	0.3396437	0.1388883	0.142111	0.12383	**0.1209646**
R^2	0.8786	0.887	0.8786	0.854	0.8666	0.8785	**0.8913236**

From Table 1, it is clear that the estimated mean square error across methods has small values except for Ridge and PCR method and it can be noted also that are no large differences between the methods according to the coefficient of determination, but the most efficient method seems to be Elasticnet.

2.2 Using 10-fold cross validation

In addition, for a more complete validation we use cross validation method "10-fold" on the data, we calculate a test error for each group and averaging, which is the estimator of the test error by cross validation. Note CV(f) this error for a given model. this method use the smallest root mean square error RMSE to find the optimal model of the 10 rounds. We obtained the following results:

Table 2. Summary table of by cross validation 10-fold.

	PCR	PLS	Ridge	LARS	Lasso	Adaptive Lasso	Elasticnet
RMSE	0.519	0.394	0.388	0.392	0.383	0.384	0.377
R^2	0.764	0.824	0.839	0.852	0.827	0.831	0.833
CV(f)	0.344	0.399	0.151	0.307	0.292	0.152	0.143

According to results the best technique is Elasticnet since it has the highest R^2 and the lowest CV(f).

References

[1] Wold, S. PLS for multivariate linear modeling. Chemometric methods in molecular design, **2**. Wiley-VCH, 1995.

[2] J. Friedman, T. Hastie, and R. Tibshirani, "Lasso and elastic-net regularized generalized linear models," Version 1.3, Jstatsoft.org, April 25, 2010.

[3] Zou and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, **67**, pp. 301-320, 2005.