Penalized regression with a combination of the L1 norm and the correlation based penalty.

Mohammed El Anbari, Abdallah Mkhadri

LA RÉGRESSION PÉNALISÉE COMBINANT LA NORME L_1 ET UNE PÉNALITÉ TENANT COMPTE DES CORRÉLATIONS ENTRE LES VARIABLES

Mohammed El Anbari & Abdallah Mkhadri

Dept de mathématiques, Batiment 425, Université Paris-Sud 11, 91405 Orsay Cedex
mohammed.elanbari@math.u-psud.fr
&
Dept de mathématiques, Laboratoire Libma, Faculté des Sciences-Semlalia, Université Cadi Ayyad, B.P. 2390 Marrakech, Maroc
mkhadri@ucam.ac.ma

RÉSUMÉ

La sélection de variables peut être difficile, en particulier dans les situations où un grand nombre de variables explicatives est disponible, avec la présence possible de corrélations élevées comme dans le cas des données d'expression génétique. Dans cette note, nous proposons une nouvelle méthode de régression linéaire pénalisée, appelée l'*elastic corr-net*, pour simultanément estimer les paramètres inconnus et sélectionner les variables importantes. De plus, elle encourage un effet de groupe: les variables fortement corrélées ont tendance à être toutes incluses ou toutes exclues du modèle. La méthode est fondée sur les moindres carrés pénalisés avec une pénalité qui, comme la pénalité L_1, rétrécit certains coefficients exactement vers zéro. En outre, cette pénalité contient un terme qui lie explicitement la force de pénalisation à la corrélation entre les variables explicatives. Pour montrer les avantages de notre approche par rapport aux méthodes les plus concurrentes, une étude sur des données simulées est réalisée en moyenne et grande dimension. Enfin, nous appliquons la méthodologie un exemple de données réelles. Si $p \gg n$, notre méthode reste compétive et elle permet aussi de sélectionner plus que n variables.

Mots-Clés: *Sélection de variables; grandes dimensions; elastic-net; effet groupement; pénalité de corrélation.*

ABSTRACT

Variable selection in linear regression can be challenging, particularly in situations where a large number of predictors is available with possibly high correlations, such as gene expression data. In this note we propose a new method called the *elastic corr-net* to simultaneously select variables and encourage a grouping effect where strongly correlated predictors tend to be in or out of the model together. The method is based on penalized least squares with a penalty function that, like the Lasso penalty, shrinks some coefficients to exactly zero. Additionally, this penalty contains a term which explicitly links strength of penalization to the correlation

between predictors. Simulation study in small and high dimensional settings is performed, which illustrates the advantages of our approach in relation to several other possible methods. Finally, we apply the methodology to real data sets. If $p \gg n$, our method remains competitive and also allows the selection of more than n variables in a new way.

Key Words: *Variable selection; high dimensional setting; elastic net; grouping effect; correlation based penalty.*

1 Introduction

Suppose that the data set has n observations with p predictors. We consider the standard linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where $\mathbf{y} \in I\!R^n$ is the response and $\mathbf{X}$ is the nxp model matrix, with $\mathbf{x}_j \in I\!R^n, j = 1, ..., p$, are the predictors. It is assumed that the response is centered and the predictors are standardized. When p is large relative to n, there are many alternative procedures that outperform the ordinary least squares (OLS) which can be categorized into one of the two groups : regularization methods (like Ridge regression) and classical variable selection. But, the final fit of Ridge regression is difficult to interpret because all p predictors will remain in the model. While, the classical variable selection is computationally heavy to implement when p is large.

Recall that the ridge regression minimizes the penalized problem:

$$\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{2}$$

here $\|.\|_2$ represents the L_2 norm and λ is a non negative tuning parameter which can be selected by cross-validation.

More recently interest has focused on an alternative class of methods which implement both the variable selection and coefficients shrinkage in a single procedure. The Lasso (Tibshirani) is a popular one for regression that uses the L_1 to achieve a sparse solution: i.e.

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}, \tag{3}$$

where $\|.\|_1$ represents the L_1 norm and λ is a non negative tuning parameter. The use of L_1 penalty on the coefficients has the effect of automatically performing variable selection by setting certain coefficient to zero and shrinking the remainder. This method was made particularly appealing by the advent of the efficient LARS algorithm (Efron et al. 2004) which compute the entire regularization path for Lasso. Although it is a highly successful technique, it has two drawbacks:

i) In $p > n$ case, the Lasso can select at most n variables, this can be a limiting feature for a variable selection method.

ii) When there are several highly correlated input variables in the data set, all relevant to the output variable, the L_1-norm penalty tends to pick only one or few of them and shrinks the rest to 0.

To improve the Lasso, Zou and Hastie (2005) proposed the elastic net, which is the solution to the penalized problem:

$$\hat{\beta}_{\text{enet}} = (1+\lambda_2)\arg\min_{\beta}\left\{\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\right\}, \tag{4}$$

where λ_1 and λ_2 are non negative tuning parameters. The elastic-net utilizes the L_1 and L_2 penalties to select variables while inducing grouping effect.

The elastic net penalty has shown improvements over Lasso in many situations. However, we observe a few limitations.

i) The elastic net penalty does not explicitly contain correlation.

ii) The elastic-net exhibits a poor performance in selecting a related variables as a group when within-group correlations are non-extreme.

In this note, we propose an alternative regularization procedure based on the penalized least squares for variable selection in linear regression problem, which combines the L_1 norm and CP (Correlation based penalty) penalties. We call it the *elastic corr-net*. The elastic corr-net performs automatic variable selection and parameter estimation where highly correlated variables are able to be selected (or removed) together. Additionally, the CP penalty contains a term which explicitly links strength of penalization to the correlation between variables.

The remainder of this note is organized as follows. Section 2, formulates the elastic corr-net as a constrained least squares problem. Computational issues, including choosing the tuning parameters, are discussed in Section 3. Finally, the elastic corr-ne is applied to simulated and real world data in medium and high-dimensional setting in Section 4.

2 The criterion

The elastic corr-net estimates solves

$$\hat{\beta}_{\text{corr-net}} = \arg\min_{\beta}\left\{\|\mathbf{y}-\mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 P_c(\beta)\right\}, \tag{5}$$

where

$$P_c(\beta) = \sum_{j=1}^{p-1}\sum_{j>i}\left\{\frac{(\beta_i-\beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i+\beta_j)^2}{1+\rho_{ij}}\right\}, \tag{6}$$

$\rho_{ij} = \mathbf{x}_i^t\mathbf{x}_j$ is the sample correlation and λ_1 and λ_2 are non negative tuning parameters. The penalty $P_c(\beta)$ was introduced by Tutz and Ulbricht (2006) as an alternative to the L_2 norm in ridge regression method.

3 Estimation and computation

It turns out that minimizing criterion (5) is equivalent to a Lasso-type optimization problem. This fact implies that the new method can enjoy the computational advantage of the Lasso: we can demonstrate that the elastic corr-net problem can be transformed into an equivalent lasso problem on augmented data. The explanatory matrix in the augmented data is a $(n+p)\times p$ matrix and it has rank p, which means that the elastic corr-net can potentially select all p predictors in all situations.

In practice, it is important to select appropriate tuning parameters λ_1 and λ_2 in order to obtain a good prediction precision. Note that there are two tuning parameters in the elastic corr-net. Typically we first pick a (relatively small) grid values for λ_2, say $(0, 0.01, 0.1, 1, 10, 100)$. Then, for each λ_2, LARS algorithm produces the entire solution path of the elastic corr-net. The other tuning parameter is selected by tenfold CV. The chosen λ_2 is the one giving the smallest CV error.

4 Examples

4.1 Grouping effect

For the illustration of the grouping effect we use de following Example: in this Example, $n = 100$ and there are 40 predictors. The true parameters are

$$\beta = (\underbrace{1.85, ..., 1.85}_{5}, \underbrace{3, ..., 3}_{5}, \underbrace{4, ..., 4}_{5}, \underbrace{0, ..., 0}_{25})^T$$

and $\sigma = 1.5$. The predictors were generated as:

$$\mathbf{x}_i = Z_1 + 0.01\varepsilon_i, \quad Z_1 \sim N(0,1), \quad i = 1, ..., 5$$

$$\mathbf{x}_i = Z_2 + 0.01\varepsilon_i, \quad Z_2 \sim N(0,1), \quad i = 6, ..., 10$$

$$\mathbf{x}_i = Z_3 + 0.01\varepsilon_i, \quad Z_3 \sim N(0,1), \quad i = 11, ..., 15$$

$$\mathbf{x}_i \sim N(0,1), \quad i = 16, ..., 40$$

where ε_i are independent identically distributed $N(0,1), i = 1, ..., 15$. In this model the three equally important groups have pairwise correlations $\rho \approx 0.99$, and there are 25 pure noise features.

4.2 Medium settings

This is a simulated example based on the model

$$\mathbf{y} = \mathbf{X}\beta + \sigma\varepsilon \tag{7}$$

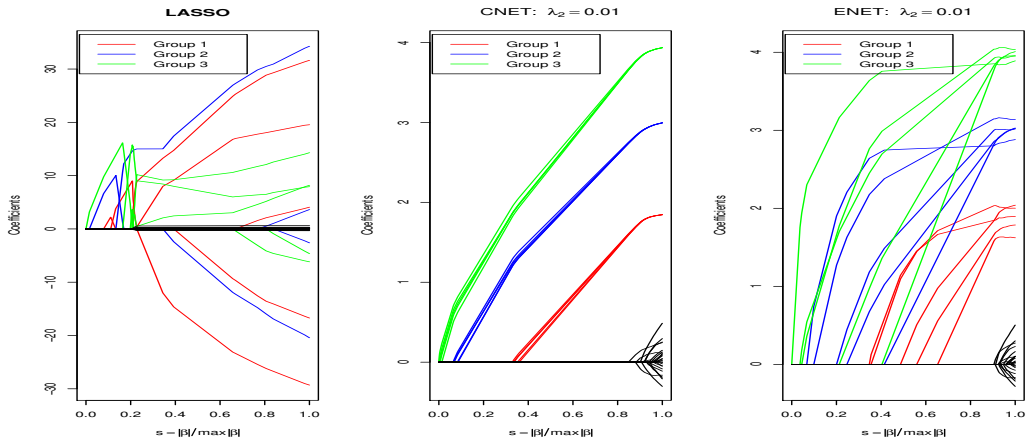

Figure 1: Solution paths of regression coefficients. From the left to right are the solution paths for Lasso (LASSO), elastic corr-net (CNET) and elastic-net (ENET): the elastic corr-net shows the "grouped selection".

where $\varepsilon \sim N(0,1)$. There are $p = 8$ predictors. The true parameters are $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma = 3$ with the correlation matrix given by $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.7^{|i-j|}$. The size of the training, validation, and test sets are $20/20/200$ respectively.

METHOD	median MSE_y	H	FP
RIDGE	4.45	3	5
LASSO	4.03	3	3
ENET	3.59	3	4
CNET	**3.52**	3	3

Table 1: The simulated example 2 - median test mean squared error MSE_y, Hits (H) and False Positives (FP) over 50 data set for different methods: the ridge (RIDGE), the Lasso (LASSO), the elastic-net (ENET) and the elastic corr-net (CNET).

4.3 High dimensional settings

This is a simulated example based on the same model (7). The size of the training, validation, and test sets are $50/50/400$ respectively. We have 50 predictors; $\beta_i = 2$ for $i < 9$ and $\beta_i = 0$ for $i \geq 9$. $\sigma = 2$ and $\text{cor}(\mathbf{x}_i, \mathbf{x}_j) = 0.9 \times 1_{i,j \leq 9}$.

METHOD	median MSE_y	H	FP
RIDGE	0.78	8	32
LASSO	0.27	8	6
ENET	0.24	8	6
CNET	**0.17**	8	6

Table 2: The simulated example 3- median test mean squared error MSE_y, Hits (H) and False Positives (FP) over 50 data set for different methods: the ridge (RIDGE), the Lasso (LASSO), the elastic-net (ENET) and the elastic corr-net (CNET).

4.4 Real world data

The body fat data set has been used by Penrose, Nelson and Fisher (1985). The study aims at the estimation of the percentage of body fat by various body circumference measurements for 252 men. The thirteen regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13).
In order to investigate the performances of the *elastic corr-net*, the data set has been split 20 times into a training set of 151 observations and a test set of 101 observations. Tuning parameter have been chosen by tenfold cross validation.

Method	median MSE_y	median no. of selected variables
RIDGE	21.02	13
LASSO	21.28	9.5
ENET	21.21	7
CNET	**20.42**	10

Table 3: Body fat data - median test mean squared error and selected variables over 20 random splits for different methods.

5 Discussion

Similar to the elastic-net method, the elastic corr-net encourages a grouping effect. Due to the efficient path algorithm (LARS), the elastic corr-net procedure enjoys the computational advantage of the elastic-net. Our simulations and empirical results have shown good performance of

our method and its superiority over its competitors in term of prediction accuracy, identification of relevant variables while encouraging a grouping effect.

Bibliographie

[1] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist., (32), 407-499*.
[2] El Anbari, M. and Mkhadri, A. (2008). Penalized regression with a combination of the L_1 norm and the correlation based penalty.*INRIA, RR-6746*.
[3] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B, 58, 267-288*.
[4] Tutz, G. and Ulbricht, J. (2006). Penalized regression with correlation based penalty. Discussion Paper 486, SFB 386, Universität München.
[5] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic-net. *J. R. Statist. Soc. B, 67, 301-320*.