Regularization algorithm and its implementation in general systems theory and its implications in physics

F Mesa¹, D M Devia¹, and G Correa¹
¹ Departamento de Matemáticas, Universidad Tecnológica de Pereira, Pereira, Colombia
E-mail: femesa@utp.edu.co

Abstract. In recent years a new physics known as complexity has emerged, whose fundamental structure is based on the general theory of systems or the so-called mathematical biology. There, problems arise with a large number of variables on which decisions must be made, and it is important to have selection techniques that are robust and efficient. In this research work, a couple of techniques are presented that will serve the aforementioned purpose; these are: the Lasso regularized linear regression technique and the Ridge regression technique, the latter with the slight difference in the penalty, since it makes use of a different rule than the Euclidean one, which implies very different consequences. One of these techniques was applied in a clinical problem and corresponds to the study of patients with diabetes. With this objective, the algorithm of descent of coordinates was implemented for the type of regularization loop and the analytical solution of regression of ridges. In addition, Kernels was used for the implementation of the regularized vector machine. The above algorithms were compared with the Ridge regularization and the Euclidean regularized vector machine. These types of techniques are common in physics, and it is vitally important to be able to implement other types of rules for a problem in which the choice of them can be simplified.

1. Introduction
The Lasso-type penalized regression (least absolute shrinkage and selection operator) is a linear regression technique proposed by Tibshirani [1], capable of selecting variables, a very important task when the number of predictors p exceeds the number of samples n. Lasso is a regularized linear regression technique, like the Ridge regression, with the slight difference in penalization, since it makes use of the p norm with p = 1 (norm ℓ₁) instead of the p norm with p = 2 (norm ℓ₂), which has important consequences.

The generic form of the regularization techniques in the context of linear models can be performed by setting an objective function which must be minimized, which depends on the observed values \( y_i \) and the estimated values \( X_{ij} \). The conditions correspond to an increasing penalty function which depends on a parameter \( \lambda \geq 0 \). A family of widely used penalty functions is the p-norm (norm \( \ell_p \)) with \( p > 0 \).

Every problem of regularization raises two important questions, the first is, ¿what is the most efficient method to minimize the proposed objective function? The second question is how to choose the most appropriate value of the adjustment parameter \( \lambda \)? From the start, the second question could be answered, which has to do with cross-validation. The answer to the first question is not so obvious, because the standard methods of regression include the diagonalization of matrices, matrix inversion or at least the solution of large systems of linear equations that are the result of having many input variables.
(predictors), becoming intractable problems. In this document the coordinate algorithm descent will be used which is a very simple algorithm, with a high stability and speed of convergence [2].

In this paper we work with two cases where \( p = 1 \) (Lasso) and \( p = 2 \) (Ridge). For \( p > 2 \) the estimator does not perform selection of variables as shown in [3]. Lasso-type regularization reduces every variable towards the origin as Ridge with the difference that some of them cancel out, this being more effective than Ridge regularization to eliminate irrelevant variables.

2. Methodology

This regression technique was initially proposed by Hoerl and Kennard [2,4] in order to eliminate the effects generated by the collinearity problem in a linear model estimated by least squares in the context \( p < n \). The Ridge regression is very similar to the least square’s regression with the difference that the coefficients are obtained by minimizing a different amount. The coefficients estimated by the Ridge methodology are the values that minimize Equation (1) [4].

\[
\hat{w}^{\text{Ridge}} = \sum_{i=1}^{n} \left( y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j^2 = \text{RSS}(w) + \lambda \sum_{j=1}^{p} w_j^2, \tag{1}
\]

where \( \lambda \geq 0 \) is the contraction parameter which must be determined separately; the Ridge method contracts the regression coefficients by including the penalty term in the objective function Equation (1), if \( \lambda \) takes a very large value, the greater the penalty and therefore the greater the contraction of the coefficients. The result of the estimation of the coefficients will be \( \hat{w}^{\text{Ridge}} \) Equation (2).

\[
\hat{w}^{\text{Ridge}}(\lambda) = (X'X + \lambda I_p)^{-1} X'Y, \tag{2}
\]

where \( I_p \) is the identity matrix of \( p \times p \). It must be considered that the coefficient \( w_0 \) is excluded from the penalty to avoid that the result depends on the origin in the variable \( Y \). Therefore, this will be estimated using \( w_0 = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \).

To avoid that the penalty varies when there are changes of scales of the variables, it is convenient to standardize these variables previously with average 0 and variance 1. Once you have estimated the coefficients \( w_j \) you should look for the value of \( \lambda \) such that, \( 0 < \lambda < \infty \) to minimize an estimate of the expected prediction error. The most common method for finding the value of the penalty parameter is cross-validation, which will be explained later. One of the shortcomings of the Ridge method is that it contracts all the coefficients towards zero, but without achieving the nullity of any of them, therefore, there is no selection of variables. Which is a very important task when you have a very large \( p \) number of predictor variables, to solve this problem we propose the Lasso regression method.

Motivated to find a linear regression technique capable of selecting variables and, of course, to stabilize estimates and predictions, Tibshirani proposed the Lasso technique [1], which is a regression technique. regularized as the one of Ridge with the difference in the term of penalty, which generates very good consequences due to the norm \( \ell_1 \). Lasso reduces the variability of the estimates by reducing the coefficients and at the same time produces interpretable models by reducing some coefficients to zero. The boom in the last few years of the Lasso technique is mainly due to the existence of a regression problem where \( p \gg n \). Lasso solves the least squares problem about the rule restriction \( \ell_1 \) Equation (3) [5].

\[
\hat{w}^{\text{Lasso}} = \sum_{i=1}^{n} \left( y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |w_j| = \text{RSS}(w) + \lambda \sum_{j=1}^{p} |w_j|, \tag{3}
\]
again, the constant $w_0$ is parameterized by standardizing the predictors and, therefore, the solution $w_0 = \bar{Y}$. In general, the models generalized by Lasso are much easier to interpret than those obtained by Ridge. As in the previous technique, you should look for the best value of the $\lambda$ parameter that minimizes the error through cross-validation [6].

The regularization techniques mentioned above depend on a penalty parameter $\lambda$, which controls the importance given to the penalty in the optimization process. The greater $\lambda$, the greater the penalty in the regression coefficients and the more costs are contracted towards zero. The choice of this parameter involves a balance between the components of bias and variance of the mean squared error (MSE) when estimating $w$. An initial proposal that continues to be suggested by some authors is the use of a Ridge trace to determine $\lambda$, consists in graphing simultaneously the estimated regression coefficients as a function of $\lambda$, and choosing the smallest value of the parameter for which these coefficients are stabilized. Another alternative for the choice of $\lambda$ is to estimate this parameter by cross-validation Algorithm 1, which consists in dividing the model into a training set to adjust a model and a test set to evaluate its predictive capacity, by means of the prediction error or other measure [7].

The way in which cross validation is applied is by dividing the set of randomly available data into $k$ subsets or folds of equal size and mutually exclusive. One of the subsets will be used as test data and the rest as training data. The cross-validation process is repeated for $k$ iterations, with each of the possible test subsets. Finally, the arithmetic average of the results of each iteration is performed to obtain a single result. The cross validation with $k = 10$ is one of the most used, but we must consider the number of observations we have available [8].

**Algorithm 1. Cross validation.**

1. The data set $D$ is divided into $k$ subsets of equal size
2. For each $i = 1, 2, ..., k$
   - The model $f^\lambda_i(z)$ is set with the training set
   - The error is calculated by cross-validation
     $$(C\text{Verror})_i^\lambda = \frac{1}{|D_i|} \sum_{z \in D_i} (y - f^\lambda_i(z))^2$$
3. Criterion: Minimize the global cross-validation error:
   $$\lambda^* = \arg \min \{C\text{Verror}\}^\lambda_i = \arg \min \left[ \frac{1}{k} \sum_{i=1}^{k} (C\text{Verror})_i^\lambda \right]$$

The coordinate descent algorithm is an optimization algorithm that is responsible for solving problems of the type of Equation (4).

$$\min h(x) := f(x) + \lambda \Omega(x), \quad (4)$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is a continuous function, which in some cases assumessuch as: $f$ is soft and convex, smooth and possibly not convex and in other cases are smooth but with restrictions in their domain. $\Omega(x)$ is a regularization function that can be non-soft and $\lambda > 0$ is the regularization parameter; $\Omega$ is frequently convex and is assumed to be separable or separable by blocks as indicated in Equation (5) [9,10].

$$\Omega(x) = \sum_{i=1}^{n} \Omega_i(x_i), \quad (5)$$

where $\Omega_i: \mathbb{R} \to \mathbb{R}$ for all values of $i$. The best example of a $\Omega(x)$ separable function is the standard $\ell^2_1$ in which $\Omega(x) = \|x\|_1 = |x_i|$. Block separability means that the identity matrix of $n \times n$ can be partitioned into vectors $U_i \forall i = 1, 2, ..., N$ such that Equation (6) [9,10].
the Lasso regression.

stable coefficient values for different values of

be the body mass index, one of the blood serum measurements and the blood pressure. There are more

Figure 3. discard those that

to determine which variables are directly or indirectly related to each other, in order to

the progress of the diabetes disease.

therefore, it can be obtained a regression with the 4 variables that provide more information to determine

patient is diabetic or not, as shown in Figure 2, which shows the relevanc

important factors for the determination of diabetes in

were optimized, obtaining

quantitative measure of the progress of the disease

average blood pressure, and six meas

4. H

2.

3.

1.

2.

3.

4.

5.

6.

7.

8.

9.

Repeat

Converge

stabilized

for the Lasso type and Ridge type regression, the regularization function can be written as Equation

\[ \Omega_\alpha(w) = \sum_{j=1}^{p} \left[ \frac{1}{2} (1-\alpha)w_j^2 + \alpha |w_j| \right], \]

(7)

where with \( \alpha = 0 \) you have the Lasso regression and with \( \alpha = 1 \) Ridge return. The coefficients \( w_j \)
using coordinate descent to solve the equation (3.1) for the Lasso and Ridge regressions, is given by Equation (8) [11].

\[ \hat{w}_j = \left( \frac{s(\frac{1}{n} \sum_{i=1}^{n} x_{ij}(y_i - \bar{y}))^{\alpha}}{1 + \lambda (1 - \alpha)} \right) \quad (8) \]

where \( y^*_j = w_0 + \sum_{l \neq j} x_{lj} \hat{w}_l \) is the adjusted value excluding the contribution of \( x_{ij} \), and \( S(z, \gamma) \) is the
soft thresholding function for the soft thresholding of the vector \( w \) which is defined as Equation (9) [12].

\[ \text{soft}(z, \gamma) = \begin{cases} 
    z - \gamma & z > 0 \quad y < |z| \\
    z + \gamma & z > 0 \quad y < |z| \\
    0 & y \geq |z|.
\end{cases} \quad (9) \]

According to the above, below, we show Algorithm 2 of coordinate descent for Lasso and Ridge.

Algorithm 2. Coordinate descent for Lasso and Ridge.

Requires: \( y, X \)

1. Standardize \( y \) and \( X \)
2. For each \( \lambda \)
3. Initialize \( w = (XX^T + \lambda I)^{-1}X^Ty \)
4. Repeat
5. for \( j = 1, \ldots, D \) do
6. \( \epsilon_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}(y_i - w^Tx_i + w_j x_{ij}) \)
7. \( w_j = \text{soft}(\epsilon_j, \lambda \alpha) \)
8. end for
9. Converge

3. Results

For this work we used a database with 442 samples, 10 variable input and 1 variable output, to determine
the progress of diabetes disease in some patients. The variables correspond to age, sex, body mass index, average
blood pressure, and six measurements of blood serum. The output variable corresponds to a
quantitative measure of the progress of the disease. With the Algorithm 2, the values of the coefficients
\( w \) were optimized, obtaining Figure 1 for the regression method of Lasso, where it is observed that some
important factors for the determination of diabetes in patients, they are due to body mass indexes and
blood pressure, while the age and sex indexes could be discriminated since they are not significant.

In addition, only some blood serum measurements could be considered for the determination if the
patient is diabetic or not, as shown in Figure 2, which shows the relevance of the input variables,
therefore, it can be obtained a regression with the 4 variables that provide more information to determine
the progress of the diabetes disease.

To validate the results obtained through Lasso regression, we calculated the correlation matrix of all
the variables to determine which variables are directly or indirectly related to each other, in order to
discard those that have any of these relationships and take into account only one of the two variables
in Figure 3.

The Figure 4 shows the graph of the coefficients \( w \), where the variables with the highest weight will
be the body mass index, one of the blood serum measurements and the blood pressure. There are more
stable coefficient values for different values of \( \lambda \), however there is no clear selection of variables within
the Lasso regression.
4. Conclusions

With the coordinate descent algorithm, it is possible to solve the optimization problem proposed in Equation (4) efficiently and quickly, for both types of Lasso and Ridge regression. In addition, with the cross-validation algorithm, the best value that the regularization parameter $\lambda$ can take can be determined. Both methods of regression, present good results respect to the values of $w$, since with the Lasso regression those variables that do not contribute much information to the output are penalized, performing the task of selecting variables. But with the Ridge regression method, you get a better value of $\lambda$ where these coefficients stabilize $w$.

These types of techniques enrich the physical and mathematical analysis since they allow the use of a non-classical norm such as the Euclidean one and allow through another type of norm to better visualize the results, we must not forget that in a finite space the norms are equivalents but the way to read and implement them are different.

The results show us that the regularization methods shown for the selection of variables, the Lasso regularization method and the Ridge regularization method offer us a great advantage if they are applied to different problems of the supposed new branch of physics called "complexity", given the immense volume of data that is handled there.
References

[1] Tibshirani R 1996 Theorem for the lasso and the ridge regression J. Royal Stat. Soc. Ser. B. 58(1) 267
[2] Hoerl A, Kennard R 1970 Ridge regression: biased estimation for nonorthogonal problems Technometrics 12(1) 55
[3] Fan J, Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle properties J. Am. Stat. Assoc. 96 1348
[4] Zhu X X, Bamler R 2010 Regularization and variable selection via the elastic net IEEE Trans. Geosci. Remote. Sens. 48(10) 3839
[5] Thompson A, Brown J C, Kay J W, Titterington D 1991 A study of methods of choosing the smoothing parameter in image restoration by regularization. IEEE Trans. Pattern. Anal. Mach. Intell. 4 326
[6] Zou H, Hastie T 2005 Regularization, and variable selection via the elastic net J. R. Stat. Soc. 67(2) 301
[7] Wu T, Lange K 2008 Coordinate descent algorithms for lasso penalized regression Ann. Appl. Stat. 2(1) 224
[8] Breheny P, Huang J 2011 Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection Ann. Appl. Stat. 5 (1) 232
[9] Simon N, Friedman J, Hastie T, Tibshirani R 2011 Regularization paths for generalized linear models via coordinate descent J. Stat. Software 39(5) 1-13
[10] Dong B, Zhang Y 2013 An efficient algorithm for \( \ell_0 \) minimization in wavelet frame based image restoration Journal of Scientific Computing 54(2) 350
[11] Fu W 1998 Penalized regressions: the bridge versus the lass J. Comput. Graphical Stat. 7(3) 397
[12] Meier L, Van De Geer S, Buhlmann P 2008 The group lasso for logistic regression J. R. Stat. Soc. Journal 70(1) 53