Regression shrinkage and selection variables via an adaptive elastic net model

Ghadeer Jasim Mohammed Mahdi, Nadia Jasim Mohammed and Zahraa Ibrahim Al-Sharea

1Department of Mathematics, College of education for Pure sciences- Ibn Al-Haitham, University of Baghdad, Iraq.
2Department of Mathematical Sciences, Fulbright College of Art and Sciences, University of Arkansas, USA.

E-mail: gmahdi@ihcoedu.uobaghdad.edu.iq

Abstract. In this paper, a new method of selection variables is presented to select some essential variables from large datasets. The new model is a modified version of the Elastic Net model. The modified Elastic Net variable selection model has been summarized in an algorithm. It is applied for Leukemia dataset that has 3051 variables (genes) and 72 samples. In reality, working with this kind of dataset is not accessible due to its large size. The modified model is compared to some standard variable selection methods. Perfect classification is achieved by applying the modified Elastic Net model because it has the best performance. All the calculations that have been done for this paper are in R program by using some existing packages.

1. Introduction

In the various fields of statistical analysis, the most commonly used model is the Multivariate Linear Regression (MLR) model. Ordinary least squares (OLS) is the most widely used estimating technique for MLR. Consider \( \{x_1, x_2, \ldots, x_p\} \) is a set of variables and \( y \) is a response, the multiple linear regression model for \( y_i, i = 1, \ldots, n \), can be written as follows:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon
\]

where \( x_{ij} \) represents the \( i^{th} \) level of the \( j^{th} \) predictor. In matrix notation, for \( n \) samples, the MLR can be expressed in the following equations:

\[
y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}
\]

where \( \varepsilon \) is the error term.
The matrix $X$ refers to the design matrix. It contains information about the predictor variable levels at which the observations are obtained. All the regression coefficients are in the vector $\hat{\beta}$, and it can be estimated using the OLS estimates; i.e., $\hat{\beta}$ can be calculated by solving the following

$$
\min_{\hat{\beta}} \{ (y - X\hat{\beta})^T (y - X\hat{\beta}) \}
$$

solving,

$$
\frac{\partial SRR(\beta)}{\partial \beta} = -2X^Ty + sX^T\beta = 0
$$

for $\beta$ gives,

$$
\hat{\beta} = (X^TX)^{-1}X^Ty
$$

(1)

The multicollinearity or collinearity exists in a linear relationship among the variables if the number of variables exceeds the number of samples [10]. The problem of multicollinearity is well explained by several studies applying different methods. Some previous methods are explained in this work, and then the modified method is compared with them.

This paper is structured as follows: In section 2, an overview for Ridge, Lasso, and Elastic Net models are given. In section 3, a modified Elastic Net model is discussed. Real data analysis is introduced in Section 4. In section 5, results and discussion are presented. Finally, the conclusion is given in section 6.

2. Ridge and Lasso and Elastic Net Models

Researchers face some challenges such as poor results and time-consuming with large datasets. One way to solve this problem is by using variable selection techniques. As a result, large datasets can be reduced to reasonable datasets. This section lists three typical variable selection (shrinkage) models which are Ridge [2], Lasso [4], and Elastic Net [11]. Shrinkage techniques can be used under certain constraints depending on the size of the datasets. The regression model can be fitted with all the variables, and the shrinkage techniques enhance the accuracy and efficiency of the model by reducing the number of variables in the model.

2.1 Ridge Model

Ridge model treats the associated variables in the same way by shrinking their coefficient similarly. Ridge estimates the coefficients of some variables to be 0 or close to 0, so these variables should be excluded from the model, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$. Similar to equation (1), Ridge seeks the vector $\hat{\beta}^{ridge}$ that minimizes the penalized RSS; i.e., $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$; hence,

$$
\hat{\beta}^{ridge} = \minimize \left( \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right), \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq K
$$

(2)

where the value of $K$ is the upper bound for the sum of the coefficients. The complexity parameter, $\lambda$, is greater than or equals to 0. If $\lambda = 0$, then $\hat{\beta}^{ridge} = \hat{\beta}$, and as $\lambda \to \infty$, $\hat{\beta}^{ridge} \to 0_p$. For $0 < \lambda < \infty$, the fitting a linear regression model and shrinkage the coefficients are balanced. The shrinkage penalty is small when $\beta_1, \beta_2, ..., \beta_p$ are close to zero [3]. Unlike OLS, Ridge solutions are not unique; as a result, before the estimation, the inputs customary can be standardized. $\beta_0$ should be estimated separately as $\tilde{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, and the remaining parameters can be estimated by using the data matrix $X$ as follows, $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^T\tilde{y}$, where $I$ is the $p \times p$ identity matrix.
2.2 Lasso Model

Another kind of penalty (LASSO) can be considered to constrain the size of OLS estimates. Lasso collapses some of the correlated parameters to 0. The $L_1$ norm is used as penalty encompasses, so

$$
\hat{\beta}_{\text{lasso}} = \min_{\beta} \left( \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right), \quad \lambda \geq 0
$$

(3)

The LASSO property allows the exclusion of variables by setting their coefficients to be 0. When some variables are excluded, the model will be more efficient, especially if the number of variables is much larger than the number of samples, $p \gg n$ [5].

2.3 Elastic Net

Some weakness appears using Lasso. For example, with highly correlated variables, Lasso sometimes does not perform well because it can select at most $n$ variables before it saturates when $p \gg n$. If there is a very high pairwise correlation in a group of variables, Lasso takes care of only one variable from the group [12]. Moreover, Lasso often performs worse than Ridge in prediction. Therefore, a penalty that combines the Lasso and Ridge constraints have been developed to overcome these limitations [2]. Elastic Net model is a variable selection procedure that uses the penalty

$$
\lambda \left[ \frac{1}{2} (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right]
$$

(4)

where $\alpha \in [0,1]$ is called the mixing parameter. Lasso and Ridge are special cases for the Elastic Net model when $\alpha = 1$ and $\alpha = 0$ as shown in equation (2) and (3). Lasso penalty gains more weight than the Ridge penalty when $\alpha$ approaches one, and the opposite happens when $\alpha$ approaches zero. Elastic Net enables us to select more than $n$ variables, improve the predictive, and tends to jointly select groups of highly correlated variables.

3. Modified Elastic Net Model

In general, the shrinkage penalty expression can be conveniently extended to $\lambda \sum_{j=1}^{p} |\beta_j|^q$, where $q > 0$. Figure 1 shows the restriction regions, $\sum_{j=1}^{p} |\beta_j|^q \leq 1$, for different values of $q$. For $q \geq 1$, the constraints are convex; and when $q$ equals 1 and 2, the lasso and Ridge penalties exist, respectively.

![Figure 1. The Contours of constant value of $\sum_j |\beta_j|^q$ for different values of $q$.](image)
To estimate the mean square error of a model, cross-validation (CV) can be used. CV is a re-sampling technique where the chosen method is applied to predict the held-out outcome. The two typical CV are, i) leave-one-out CV: it takes out one observation, then on the remaining \((n - 1)\) observations the model is fitted; ii) \(k\)-fold CV: it splits the data randomly into \(k\) equal size subsets (folds), then eliminates one, and fits the model on the remaining \((k - 1)\) folds. In both leave-one-out CV and \(k\)-fold CV, the process is repeated \((n - 1)\) and \((k - 1)\) times; respectively, then we take the average. Leave-one-out CV has low bias and high variance; in contrast, \(k\)-fold CV has adequate bias and low variance. To minimize the estimated \(\sum\), \(k\)-fold CV is used to choose the value of \(\lambda\). Typically, the value of \(k\) between 5 and 10 is indicated as good for computational.

The weight of the parameters is the key part of the modified Elastic Net model. It could be estimated by

\[
\hat{\lambda}_k = \left(\hat{\beta}_k^{\text{inp}}\right)^{-\gamma},
\]

where \(\gamma\) is a positive constant and \(\text{inp}\) is \(\hat{\beta}_k^{\text{inp}}\) is a set of the initial parameters that could be selected using a naïve Elastic Net model. Therefore, the modified Elastic Net model penalty can be written as:

\[
\lambda \left[ \frac{1}{2} (1 - \alpha) \sum_{k=1}^{p} \beta_k^2 + \alpha \sum_{k=1}^{p} |\beta_k| \right] + \alpha^* \sum_{k=1}^{p} \hat{\lambda}_k |\beta_k| \tag{5}
\]

The adaptive weight \(\hat{\lambda}_k^{(s)} = \frac{1}{\left|\hat{\beta}^{(s-1)}(\alpha^{(s-1)})\right|} j = 1, 2, ..., p\). For \(s = 1\), the adaptive weights are equaled to the naïve Elastic Net weights. When \(s = 2, ..., p\), \(\hat{\lambda}_k^{(s)}\) corresponds to the adaptive weights for the previous step. Figure 2 provides a simple geometrical view of Lasso and modified Elastic Net models. The solid gray areas are the regions for both Lasso and modified Elastic Net models. The red ellipses correspond to the contours of the residual sum of squares functions. \(\hat{\beta}\) depicts the normal (unconstrained) least-squares estimate. The curve contours encourage strongly correlated variables to share coefficients. The plot indicates that the modified Elastic Net yields less 0-valued estimates than lasso estimation [8]. It is an attractive feature, especially when \(p \gg n\). The following algorithm summarizes the modified Elastic Net model approach.

**Algorithm: Modified Elastic Net Model**

The choice of \(\alpha\) and \(\lambda\) in the Elastic Net model.

In practice, a grid of \(\alpha\), \(\{\alpha_1, \alpha_2, ..., \alpha_M\}\), can be constructed and for each \(m = 1, 2, ..., M\):

- For given \(\alpha = \alpha_m\), \(k\)-fold CV \(\lambda\).
- For \(m = 1, ..., M\), save the Mean Squared Error.
- Compare the M Mean Squared Error.
- Choose \(\alpha\) that is associated with the preferred value.
- Use the best \(\lambda\) within the selected profile for modeling.
4. Real data Analysis

Diagnosis and classification of cancer are complicated due to the broad datasets that deal with this type of research. One of the universal forms of cancer is the Leukemia dataset. Dataset was published by Golub in 1999 [7]. The complete Golub-Merge dataset is available in the golub Esets packages in R program. Working with this dataset confront many difficulties such as time-consuming and inefficient performance, as the number of predictors exceeds the number of responses. Leukemia dataset contains 3051 variables (genes) and 72 samples (patients). It is used in this section for experimental evaluation. It shows how the gene expression monitoring (via the DNA microarray) can identify the new cancer cases, providing a standard method for assigning tumors to known classes [9]. Patients can be categorized into two types, which are: Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) [6].

4.1 Analyzing Golub Datasets

Significant genes for the type of cancer must be selected to obtain a reduced dataset. This means that only specific genes that are expressed differently across classes should be advised. As the differential genes are defined between two groups, the t-test appears to be a common choice [1] [13]. However, the t-test requires a normality assumption, which may not be a logical assumption. A histogram of 3051 × 2 cannot be plotted to have an idea about the justification for normality. Mann Whitney U test looks more suitable and is almost as efficient as the t-test. After running the t-test, the p-values are modified according to the Benjamin-Hochberg method. Most genes do not appear to have different mean values across classes. The same result for the median has been achieved. The mean and median differences between types are shown in figure 3. Most of the genes are clustered around zero. A few numbers of genes seem interesting, and they can be easily indicated.
5. Result and Discussion for the Real Data

The results of OLS, Ridge, Lasso, Elastic Net, and modified Elastic Net results are shown in Table (1). The accuracy rate of classification, the number of selected genes, and the approximate time-consumption are determined for all methods. The best accuracy occurs with the modified Elastic Net model with only 63 selected genes and approximately 9 seconds of consuming time. The paths and the misclassification errors for both Elastic Net and modified Elastic Net models are displayed in figure 6. Figure 6.a shows that the Elastic Net model has selected 329 genes as significant variables. Figure 6.b shows that the modified Elastic Net model has selected only 63 genes out of the total of 3051 genes. In both models, the dotted vertical lines indicate the optimal value of $\lambda$. It is “3.36” and “-4.47” in both in Elastic Net and modified Elastic Net models, respectively. The factor map is plotted with two dimensions in figure 7. The percentage of variation for both first and second dimensions are 50.34 and 7.98, respectively. The selected genes in the modified Elastic Net model are used to construct the model. The training dataset was classified by type of cancer samples. The model was applied to the testing dataset as well. Figure 7.b shows the ALL in black and AML in red colors. Only one sample (ID=66) is misclassified, has an AML class, but it is classified as ALL.

Table 1. Comparison between OLS, Ridge, Lasso, Elastic Net, and modified Elastic Net for classification Leukemia dataset.

| Methods            | Accuracy rate (%) | Number of selected genes | Approximately consuming time |
|--------------------|-------------------|--------------------------|------------------------------|
| OLS                | 56.93             | 3051                     | ≈ 30 seconds                |
| Ridge              | 83.00             | 3051                     | ≈ 24 seconds                |
| Lasso              | 85.33             | 23                       | ≈ 12 seconds                |
| Elastic Net        | 88.93             | 329                      | ≈ 18 seconds                |
| Modified Elastic Net | **95.80**         | 63                       | ≈ 9 seconds                 |
Figure 4. a. Coefficients and Misclassification Error for Elastic Net Model

Figure 4. b. Coefficients and Misclassification Error for the modified Elastic Net Model
Figure 5 a. Factor map with two dimensions.

Figure 5 b. Classification plot for patients according to AML and ALL classes.

Figure 5. Factor map and classification plot for Leukemia dataset.
6. Conclusions
In this paper, an enhanced Elastic Net model was developed and applied to the leukemia dataset. It is derived by combining Lasso and Ridge models with the addition of a penalty term. The modified Elastic Net is compared with some standard classification methods. The updated Elastic Net is the best variable selection model in both training and testing datasets because it has high accuracy results and less time-consuming. The classification accuracy using the modified Elastic Net model is 95.80%, while the other is below 89%. Among 3051 genes, 63 genes have been selected by the updated Elastic Net model. Approximately 9 seconds was required to complete the classification process. The proposed model can be used for large datasets in the future because it is accurate and efficient.

References
[1] Arpino G, D Generali, A Sapino, L Del Matro, A Frassoldati, M de Laurentis, P Pronzato, G Mustacchi, M Cazzaniga and S De Placido 2013 Gene expression profiling in breast cancer: a clinical perspective *The Breast Cancer* 22(2) pp 109–120
[2] D Aydin, B Yüzbüş and S E Ahmed 2016 Modified ridge type estimator in partially linear regression models and numerical comparisons *Journal of Computational and Theoretical Nanoscience* 13(10) pp 7040-7053
[3] E Bair, T Hastie, D Paul and R Tibshirani 2006 Prediction by supervised principal components *Journal of the American Statistical Association* 101(473) pp 119–137
[4] H Choi, E Song, S S Hwang and W Lee 2018 A modified generalized lasso algorithm to detect local spatial clusters for count data *ASA Advances in Statistical Analysis* 102(4) pp 537-563
[5] G J Mahdi 2020 A Modified Support Vector Machine Classifiers Using Stochastic Gradient Descent with Application to Leukemia Cancer Type Dataset *BSJ* 17(4) pp 1255-1266 DOI: 10.21123/bsj.2020.17.4.1255
[6] Q Zhang, G Mahdi, J Tinker and H Chen 2020 A graph-based multi-sample test for identifying pathways associated with cancer progression *Computational Biology and Chemistry* 26 107285 DOI:10.1016/j.compbiolchem.2020.107285
[7] T R Golub, D K Slonim, P Tamayo, C Hayward, M Gaasenbeek, J P Mesirov and C D Bloomfield 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring *Science* 286(5439) pp 531-537
[8] G J Mahdi 2018 *Hierarchical Bayesian Regression with Application in Spatial Modeling and Outlier Detection Diss* University of Arkansas, Fayetteville
[9] J Li, W Dong, D Meng and H Xiao 2016 Gene selection for cancer classification using improved group lasso *Chinese Control and Decision Conference (CCDC)* IEEE pp 4221-4225
[10] A F Lukman, K Ayinde, S Binuomote and O A Clement 2019 Modified ridge-type estimator to combat multicollinearity: *Application to chemical data Journal of Chemometrics* 33(5) pp e3125
[11] L T P Thao and R Geskus 2019 A comparison of model selection methods for prediction in the presence of multiply imputed data *Biometrical Journal* 61(2) pp 343–356 https://doi.org/10.1002/bimj.201700232
[12] G J Mahdi, A Chakraborty, M E Arnold and A G Rebelo 2019 Efficient Bayesian modeling of large lattice data using spectral properties of Laplacian matrix *Spatial Statistics* 29 pp 329–350
[13] G J Mahdi, B A Kalaf and M A Khaleel 2020 Enhanced Supervised Principal Component Analysis for Cancer Classification *IJIS* 62(4) Accepted paper
[14] B A Hameed, A N Salman, B A Kalaf 2020 On Estimation of P (Y_1< X< Y_2) in Cased Inverse Kumaraswamy Distribution *Iraqi Journal of Science* 26 pp 845-53