May 2025

Performance of LASSO and Ridge Regression for Variable Selection in Genome-Wide Association Studies of Maize Flowering Time

Dipok Deb
PhD Student, Big Data Analytics, UCF, di876893@ucf.edu

Performance of LASSO and Ridge Regression for Variable Selection in Genome-Wide Association Studies of Maize Flowering Time

Dipok Deb
Data Sciene I (STA 6366)
Department of Statistics and Data Science
University of Central Florida
Email: dipok.deb@ucf.edu

***Abstract*—Genome-Wide Association Studies (GWAS) are instrumental in identifying genetic variants linked to complex traits, providing valuable insights into trait heritability and biological mechanisms. This study applies GWAS to investigate flowering time in maize, a critical adaptive trait, using a diverse dataset of 5,000 recombinant inbred lines across eight environments. Traditional GWAS methods often encounter challenges in high-dimensional datasets due to the presence of multiple small-effect genetic loci. To address this, we compared two penalized regression methods—LASSO and Ridge regression—to perform variable selection and regression analysis within a GWAS framework. LASSO effectively reduced the number of predictors by selecting the most impactful variables, while Ridge regression retained more features, offering a broader genetic context for predicting flowering time. Results demonstrated that Ridge regression yielded slightly better predictive performance, achieving a lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) than LASSO.**

***Index Terms*—Machine Learning, Genome-Wide Association Study (GWAS), LASSO Regression, Ridge Regression, Variable Selection, Predictive Modeling in Genomics.**

I. Introduction

A Genome-Wide Association Study (GWAS) is a powerful technique for uncovering genetic variants linked to specific traits or diseases within populations [1]. By analyzing the genomes of large groups of individuals, GWAS identifies common single nucleotide polymorphisms (SNPs) or genetic loci that differ between individuals with and without a particular condition. Typically, GWAS involves collecting DNA samples, genotyping a vast array of SNPs across the genome, and statistically assessing the association between each SNP and the trait of interest. This approach is particularly valuable for studying complex traits, which often involve multiple genetic variants with small effects rather than single large-effect loci.

One such complex trait is flowering time, a critical adaptive characteristic in plants that enables them to synchronize their lifecycle with environmental conditions. Buckler et al. [2] investigated flowering time in the cross-pollinating species Zea mays (maize) by leveraging a GWAS approach across 5,000 recombinant inbred lines (NAM population) in eight distinct environments. Unlike self-pollinating species like rice and Arabidopsis, where large-effect QTLs for flowering time are common, maize displayed numerous small-effect QTLs shared across families, with varying allelic effects across founder lines. No single QTLs were found to correlate strongly with factors such as geographic origin, epistasis, or environmental interactions These findings suggest that flowering time in maize follows a simple additive genetic model, providing insight into its genetic architecture and the ways in which GWAS can reveal the subtle contributions of many genetic variants to complex traits.

To enhance GWAS analyses of complex traits like flowering time, researchers are increasingly turning to advanced statistical methods. For example, Waldmann et al. [3] explores the effectiveness of penalized regression techniques, specifically the Lasso and Elastic Net methods, in GWAS. Their study compares these approaches using simulated datasets with varying levels of correlation between SNPs, as well as real cattle data focusing on milk fat content. These penalized regression methods help manage the high dimensionality and correlation within GWAS datasets, identifying significant genetic markers associated with traits.

The aim of this study is to compare statistical models, specifically Lasso and Ridge regression, for performing variable selection and regression analysis within a typical genome-wide association (GWA) framework using a maize dataset. The remaining sections are organized as follows: Section II discusses the methodology, Section III outlines data analysis, and Section IV presents the conclusions and limitations of this study.

II. Methodology

In this project, we employ Lasso Regression and Ridge Regression to identify the key features in the maize dataset. Subsequently, we apply a penalized multiple linear regression model using the variables selected from Lasso and Ridge Regression. Finally, we assess the model's performance by calculating the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

A. Dataset

In this work we used Maize data [2]. The dataset includes 25 crosses, also referred to as families or populations, each consisting of roughly 200 recombinant inbred lines (RILs). It contains measurements for various phenotypes along with genetic marker data. In this dataset there are 4,981 observations across 7,393 variables, with some missing values. The primary phenotype studied here is the time to male flowering (dtoa), while 7,389 independent variables represent the SNP markers.

B. Methods

1) Lasso Regression: Lasso regression, Least Absolute Shrinkage and Selection Operator, is a linear regression method that uses **L1 regularization** to minimize overfitting and facilitate feature selection. It adds a penalty term, proportional to the absolute values of the coefficients, to the ordinary least squares (OLS) objective function. This **L1** penalty reduces certain coefficients to exactly zero, effectively eliminating them from the model and resulting in a sparse solution.

The objective function for Lasso regression is given by:

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where:

- y_i is the observed response for the i-th sample,
- X_i is the vector of features for the i-th sample,
- β is the vector of model coefficients,
- p is the total number of features,
- $\lambda \geq 0$ is the regularization parameter, which controls the degree of shrinkage.

As λ increases, more coefficients are set to zero, leading to a sparser model. Lasso is particularly useful when there are many features, as it performs both regularization and feature selection.

2) Ridge Regression: Ridge regression is a regularized linear regression approach that uses an **L2 penalty** (the squared sum of the coefficients) added to the OLS objective function. Unlike Lasso regression, Ridge does not zero out coefficients but instead shrinks them closer to zero, thereby lessening the impact of less significant features while keeping all features in the model.

The objective function for Ridge regression is expressed as:

$$\min_{\beta} \left(\sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

where:

- The terms are defined as in Lasso,
- $\lambda \geq 0$ controls the degree of shrinkage.

With larger values of λ, the coefficients are more strongly penalized, leading to smaller values but without setting them to zero.

C. Evaluation Metrics

We will assess the model's performance using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

III. Data Analysis

A. Dataset Exploration

In the maize dataset, there are 4981 samples, each characterized by 7393 features. The feature *Geno_Code* is a categorical variable, while the remaining features consist of numerical data. The *Geno_Code* feature is converted into numerical data using a label encoder. Our dataset contained 487 samples with missing values. These samples is removed to ensure data completeness. After discarding the samples with missing values, the dataset size is reduced to (4494, 7393). Our dependent variable is *dtoa*. To better understand its distribution, we conducted a descriptive statistical analysis. Table I provides a summary of the key descriptive statistics for the *dtoa* variable.

We split the dataset into training and test sets using an 80-20 ratio. This resulted in 3,595 samples for training and 899 samples for testing.

TABLE I
DESCRIPTIVE STATISTICS OF THE "DTOA" VARIABLE

Statistic	Value
Count	4494.000000
Mean	77.152452
Standard Deviation (Std)	3.673233
Minimum (Min)	66.015800
25% Percentile	74.847850
Median (50% Percentile)	77.343250
75% Percentile	79.401100
Maximum (Max)	91.234900

B. Model Training

We trained the LASSO and Ridge Regression models using the `scikit-learn` library in Python. In `scikit-learn`, the parameter λ for LASSO and Ridge is represented as α. For this project, we will use α as the tuning parameter for both LASSO and Ridge Regression instead of λ.

C. Result and Discussion

In this section, we present the results of both models and analyze their performance.

1) LASSO: In LASSO, we tested α values ranging from 0.1, 0.2 up to 20 to determine the optimal α using 5 fold cross-validation with the `GridSearchCV` method. Fig. 1 illustrates the mean cross-validation MSE against the logarithmic value of the regularization parameter α in a LASSO model. The curve demonstrates that the MSE initially decreases, reaching its minimum near $\log(\alpha) = -1.0$, and then rises as α increases. The optimal α is found at 0.1, marked by a dashed vertical line, where the MSE is minimized. The best α value found in our case was 0.1. With $\alpha = 0.1$, LASSO selected the top 45 features, and their corresponding coefficients are presented in Table II.

Next, we ran a LASSO regression model using these variables and the optimal alpha parameter, then calculated the MSE and RMSE as shown in Table III.

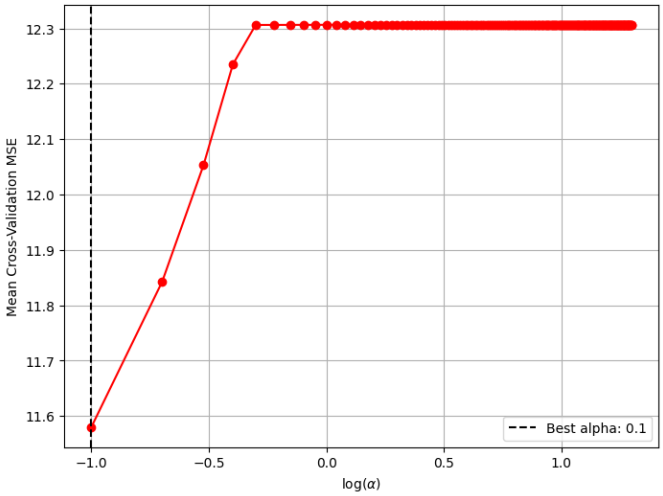

Fig. 1. Graph of Mean Squared Error (MSE) for selecting the optimal α value using 5-fold cross-validation with LASSO.

2) Ridge Regression: Ridge regression does not eliminate coefficients entirely; rather, it gradually shrinks them, enabling all predictors to stay in the model, though with a diminished influence. In our project, we used cross-validation with `GridSearchCV` to determine the optimal α value, testing a range from 10^{-3} to 10^{3}. Figure 2 illustrates how the MSE varies with different α values in Ridge regression when cross-validation is applied. As α increases from 10^{-3} to 10^{3}, the MSE initially remains high, indicating overfitting at very low α values. When α reaches values around 10, the MSE significantly decreases, showing the model achieving better generalization. Beyond this point, the MSE continues to decline gradually, suggesting that higher α values contribute to reduced error in the model. After identifying the best α, we applied a threshold value to filter out variables with minimal impact, setting it at 0.01. This threshold allowed us to remove features with coefficient absolute values below 0.01, as they had negligible contributions to the model. With an optimal α of 1,000, we implemented this filtering criterion, reducing the feature count from 7,393 to 4,211.

Using the selected features and optimal alpha parameter, we next applied a ridge regression model. The resulting MSE and RMSE values for this model are shown in Table III.

The Table III shows a comparison between Lasso and Ridge regression models. Lasso selected a smaller set of features (45) with an α of 0.1, resulting in an MSE of 12.7045 and RMSE of 3.5643. In contrast, Ridge regression retained more features (4,211) with a higher α of 1000, achieving a lower MSE of 12.2223 and RMSE of 3.4960, indicating slightly better predictive performance. Additionally, Ridge retained more features, which may contribute to capturing more information from the data.

TABLE II
SELECTED FEATURES WITH LASSO COEFFICIENTS

Serial Number	Feature	Lasso Coefficient
1	m248	0.022439
2	m266	0.028907
3	m290	0.099707
4	m355	0.103343
5	m372	0.027836
6	m439	0.178108
7	m454	0.068795
8	m1363	0.039507
9	m1424	0.004726
10	m1435	0.068852
11	m1602	0.017206
12	m1637	0.070674
13	m1652	0.000921
14	m1732	0.054523
15	m1780	0.068207
16	m2153	0.055822
17	m2163	0.008814
18	m2358	0.079268
19	m2419	0.037049
20	m2426	0.021562
21	m2434	0.093896
22	m2502	0.017274
23	m2505	0.027607
24	m2572	0.064785
25	m2967	0.096016
26	m4071	0.020212
27	m4709	0.023929
28	m4772	0.056551
29	m4860	0.091545
30	m4947	0.019050
31	m5807	0.090463
32	m5808	0.010417
33	m5815	0.050908
34	m5860	0.079617
35	m5925	0.025109
36	m5927	0.032241
37	m6369	0.001443
38	m6463	0.116776
39	m6493	0.207368
40	m6513	0.072920
41	m6559	0.006808
42	m6583	0.002906
43	m6863	0.015920
44	m6990	0.049813
45	Entry	0.001109

TABLE III
COMPARISON OF LASSO AND RIDGE REGRESSION

Method	Selected Features	α	MSE	RMSE
Lasso	45	0.1	12.7045	3.5643
Ridge	4211	1000	12.2223	3.4960

IV. CONCLUSION AND LIMITATION

In this project, we evaluated the effectiveness of two penalized regression models, LASSO and Ridge regression, for variable selection and regression analysis using maize data. Our results indicated that Ridge regression produced a better model, as evidenced by lower MSE and RMSE values. In contrast, with the LASSO model, we encountered limitations when testing alpha values below 0.1, as the model failed to converge for any α less than 0.1. The optimal parameter we obtained was $\alpha = 0.1$, but we could not explore smaller values to improve model performance further. Addressing this

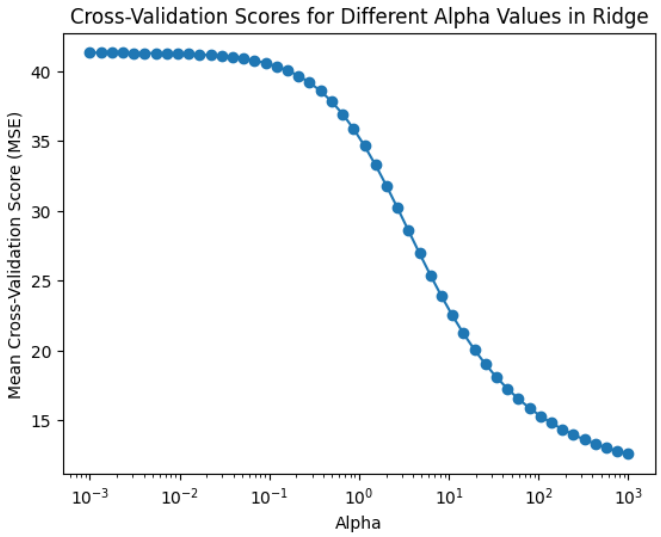

Fig. 2. Mean Squared Error (MSE) for α value using 5-fold cross-validation with Ridge.

limitation will be a focus of future work.

REFERENCES

[1] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma, "Genome-wide association studies," Nature Reviews Methods Primers, vol. 1, p. 59, 2021. doi: 10.1038/s43586-021-00056-9.

[2] E. S. Buckler, J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz, *et al.*, "The genetic architecture of maize flowering time," Science, vol. 325, no. 5941, pp. 714–718, 2009.

[3] P. Waldmann, G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner, "Evaluation of the lasso and the elastic net in genome-wide association studies," Frontiers in Genetics, vol. 4, p. 270, 2013.