A Comparative Study of Penalized Regression and Machine Learning Algorithms in High Dimensional Scenarios

Gabriel Ackall[1]*, Connor Shrader[2]*, Seongtae Kim[3†]

[1]Georgia Institute of Technology, Civil Engineering, Atlanta, GA

[2]University of Central Florida, Mathematics, Orlando, FL

[3]North Carolina A&T State University, Mathematics and Statistics, Greensboro, NC

*Authors Contributed Equally

[†]Correspondence: skim@ncat.edu

December 13, 2021

Abstract

With the prevalence of big data in recent years, the importance of modeling high dimensional data and selecting important features has increased greatly. High dimensional data is common in many fields such as genome decoding, rare disease identification, economic modeling, and environmental modeling. However, most traditional regression machine learning models are not designed to handle high dimensional data or conduct variable selection. Penalized linear regression models are a type of model designed to handle this type of data. In this paper, we investigate the use of penalized regression methods such as ridge, least absolute shrinkage and selection operation, elastic net, smoothly clipped absolute deviation, and minimax concave penalty compared to traditional machine learning models such as random forest, XGBoost, and support vector machines. We compare these models using factorial design methods for Monte Carlo simulations in 540 environments, with factors being the number of predictors, number of samples, signal to noise ratio, covariance matrix, and correlation strength. We also compare different models using empirical data to evaluate their viability in real-world scenarios. We evaluate the models using the training and test mean squared error, variable selection accuracy, β-sensitivity, and β-specificity. From our investigation, our findings indicate that the performance of penalized regression models is comparable with traditional machine learning algorithms in most high-dimensional situations or in situations with a low number of data observations. The analysis helps to create a greater understanding of the strengths and weaknesses of each model type and provide a reference for other researchers on which machine learning techniques they should use, depending on a range of factors and data environments. Our study shows that penalized regression techniques should be included in predictive modelers' toolbox.

Keywords: penalized regression, variable selection, classification, machine learning, large p small n problem, Monte Carlo simulations

1 Introduction

In the modern world, machine learning techniques such as random forest, gradient boosting, and support vector machines are often touted as versatile one-size-fits-all solutions when it comes to modeling big data [28]. This is due in part to tree based models such as XGBoost winning numerous machine learning competitions [28]. While this versatility is frequently the case, an increasingly common type of data set where there are more predictors than observations can pose challenges for these machine learning algorithms. In these situations, lesser known statistical modeling techniques that perform variable selection can potentially perform equivalently or even better than these machine learning techniques. However, there is a distinct lack of academia focusing on comparing these variable selection techniques with the more traditional machine learning techniques. This paper serves to help bridge that gap.

In these situations where there are more predictors, p, than observations, n, many traditional machine learning techniques either become infeasible to use or fail to give good predictions. The large number of predictors and small number of observations make it easy for such models to *overfit*, meaning that the models become fine tuned to the exact training data; instead of finding generalized patterns for a population of data, they memorize specific occurrences in the training data [21, 16]. Because of this, overfitted models are sensitive to new data which causes them to perform extremely well on the training data, but poorly on testing data or when deployed in the real world. Because a model's predictions in real world scenarios and on new data is the entire purpose of a model, it is very important to reduce overfitting so that predictive accuracy in these scenarios is maximized.

This paper investigates several methods to handle high-dimensional data, including the large p, small n problem. *Variable selection* techniques overcome this issue by using only a subset of the available predictors. There are many ways to implement variable selection in models. First, we considered wrapper methods such as forward selection, backward selection, stepwise forward selection and stepwise backward selection using both Akaike information criterion (AIC) and Bayesian information criterion (BIC) as the stopping criteria for the models [1, 32]. These models fit several linear models using different subsets of predictors and selects the model that optimizes either the AIC or BIC. In addition, we studied penalized regression models such as ridge regression [20], least absolute shrinkage and selection operation (lasso) [33], elastic-net (e-net) [43], smoothly clipped absolute deviation (SCAD) [14], and minimax concave penalty (MCP) [40]. These models simultaneously select important predictors and fit a linear model. Models that perform variable selection are suitable for applications such as genomics, where there are hundreds or thousands of predictors; see, for example, [34, 23].

We also evaluated the performance of several machine learning models: random forests (RF) [5], gradient boosting in the form of XGBoost [7], and support vector machine (SVM) models [8]. These types of models do not assume a linear relationship between a response and its predictors. This allows the machine learning models models to have better predictive performance on data sets where the relationship between the response and its predictors is non-linear; on the other hand, this also makes the machine learning models more susceptible to overfitting. In many applications, a combination of machine learning and variable selection is employed [36].

To compare these different techniques, models were trained and evaluated using both

Monte Carlo simulations and empirical genomic data. We are particularly in understanding in the predictive performance of these models, so we evaluated the models using the mean squared error (MSE) on both training and test data. For the linear models fitted on simulated data, we also measured the β-sensitivity and β-specificity metrics, which evaluate the ability for these models to identify important predictors [25].

Section 2 contains details about each model and details the implementation of these models for our study. Section 3 describes our simulation study design and results, while section 4 explains our empirical data analysis and results. Section 5 is a discussion of our results and Section 6 is the conclusion.

2 Methodology

2.1 Modeling Background

Suppose that we have p predictor variables $X_1, X_2, \ldots, X_p$ and one response variable Y that depends on some (or all) of the predictors. We assume that Y can be expressed as

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \tag{1}$$

where f is a function and ϵ is an independent random error with mean zero. The goal of supervised modeling is to find a function $\hat{f}$ that is a suitable approximation for f using a *training set*. This study focuses on *regression modeling*, where the response Y is a number on a continuous interval.

In practice, the function f that relates the predictors to the response is complex. Most statistical models assume that f takes some particular form and estimates a function $\hat{f}$ of that form. For example, many regression models assume that f is a linear function of the predictors; that is, linear models assume that

$$f(X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2}$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are coefficients that the models attempt to estimate.

The most common method to estimate the coefficients in a linear model is with *ordinary least squares* (OLS), which selects the values $\beta_0, \beta_1, \ldots, \beta_p$ that minimize the residual sum of squares

$$\text{RSS} = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}) \right]^2 \tag{3}$$

OLS is common because it is the best linear unbiased estimator; that is, OLS has a lower variance than any other linear unbiased estimator [18, 16]. However, if the number of predictors p is large compared to the number of observations n, OLS will overfit to the training data. Furthermore, if p exceeds n, then the OLS has infinitely many solutions that simply interpolate the training data. In these cases, OLS becomes unreliable for making predictions on test data.

Other types of linear models can overcome this large p small n problem by introducing a small amount of bias. In many cases, these models can perform *variable selection*

by setting the coefficients of unimportant predictors to zero. There are several ways to implement variable selection into a linear model. *Filter methods* work by evaluating the ability for each individual predictors to predict the response; then, a model is fit using the predictors selected [30, 11]. *Wrapper methods* fit models using different subsets of predictors and choose the model that has the best performance [19, 25]. Finally, *embedded methods* perform variable selection during the model training process [19, 25]. This paper focuses on wrapper methods and embedded methods. In addition, we considered several non-linear machine learning methods to draw a comparison between linear regression models and machine learning models.

2.2 Subset Selection Methods

Subset selection methods are wrapper methods that attempt to find a subset of the predictors $X_1, X_2, \ldots, X_p$ that are most correlated with the response variable Y. These algorithms usually fit models for many different subsets and choose the subset of predictors that results in the best model. Although subset selection techniques can be applied to many types of models, we will focus on subset selection with linear regression.

There are two main benefits to using subset selection methods. By reducing the set of available predictors to just those that are strongly related to the response, overfitting can be mitigated. Another benefit of subset selection is that it creates a more interpretable model. If a data set includes thousands of predictors but only a few are related to the response, a model found using subset selection will be easier to understand than a model that relies on all of the parameters.

Best subset selection is a wrapper method that fits and evaluates models using every possible subset of predictors. Although this method is guarenteed to find the optimal model for some evaluation metric, it is computationally infeasible with a large number of predictors [27]. In many cases, using more greedy algorithms can lead to comparable results.

Forward stepwise selection starts by fitting a model with none of the predictors (by simply estimating each observation to be the mean of the response). The algorithm then iteratively chooses the predictor that best increases the model fit until a stopping condition is met. *Backward stepwise selection* does the same thing, but starts with a full model and works backwards. In addition, *forward stepwise selection* and *backward stepwise selection* are hybrid techniques that can both add and remove predictors in each iteration. Note that backward selection and backward stepwise selection can only be used when $p < n$, since they rely on starting on a full OLS model.

Despite improving the computational costs associated with best subset selection, these alternative algorithms must still fit a massive number of models for large values of p. Consequently, the embedded methods discussed next are often favored.

2.3 Penalized Regression

In general, *penalized regression* works by fitting a model that punishes large coefficient estimates. By forcing coefficient values to shrink, the resulting model will have relatively low variance. Most, but not all, of these methods can perform variable selection during the

fitting process, making them a type of embedded method.

Almost all of the penalized regression methods in this paper solve an optimization problem of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \Big[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \Big]^2 + \sum_{j=1}^{p} P(\beta_j) \right\} \tag{4}$$

where the first summation is the usual residual sum of squares and $P(\beta)$ is a penalty function that is applied to each coefficient (not including the intercept β_0). This penalty usually depends on at least one tuning parameter (commonly denoted by λ) that controls how strong the penalty is. A suitable choice for the tuning parameter(s) will lead to a well-performing model.

Ridge regression is a penalized linear regression model that uses the penalty function $P(\beta) = \lambda\beta^2$, where $\lambda > 0$ is a tuning parameter [20]. Ridge regression benefits from having a closed-form solution that is easy to compute; it is also known for its ability to handle colinearity. However, unlike other models in this section, ridge regression is unable to perform variable selection. Another disadvantage of ridge regression is that it is a biased model, meaning that the expected values of the coefficient estimates will differ from the true coefficient values.

The *least absolute shrinkage and selection operation* (lasso) is a shrinkage method with a very similar form to ridge regression [33, 21]. The penalty function for lasso is $P(\beta) = \lambda|\beta|$, where $\lambda > 0$. Like ridge regression, the lasso is a biased estimator. One significant advantage of lasso is that it can perform variable selection by setting coefficient estimates to zero.

Elastic-net (E-net) regression use the penalties of both ridge regression and the lasso [43]. Its penalty function is $P(\beta) = \lambda_1|\beta| + \lambda_2\beta^2$, where $\lambda_1, \lambda_2 > 0$ are separate tuning parameters. An equivalent way to express this penalty function is with $P(\beta) = \lambda((1-\alpha)\beta^2 + \alpha|\beta|)$, where $\lambda > 0$ and $\alpha \in [0, 1]$ are tuning parameters. Note that if $\alpha = 0$, then the resulting model is just ridge regression, while using $\alpha = 1$ gives the lasso. The resulting model gains the advantages of both ridge regression and the lasso, but also suffers from their disadvantages, namely that they are both biased models.

The last two penalized linear models that we considered are *Smoothly-Clipped Absolute Deviation* (SCAD) and *Minimax Concave Penalty* (MCP) [14, 37, 40]. SCAD uses the penalty function

$$P(\beta) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda \\ \frac{2a\lambda|\beta| - \beta^2 - \lambda^2}{(a-1)}, & \lambda \le |\beta| < a\lambda \\ \frac{\lambda^2(a+1)}{2}, & a\lambda < |\beta| \end{cases} \tag{5}$$

while MCP uses

$$P(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2a}, & |\beta| \le a\lambda \\ \frac{1}{2}a\lambda^2, & a\lambda < |\beta| \end{cases} \tag{6}$$

These methods use piecewise penalty functions that punish larger coefficients less severely. The resulting models are consequently less biased. Another feature of SCAD and MCP is their oracle-like properties [14, 40]. This means that as $n \to \infty$, SCAD and MCP will correctly identify exactly which predictors should have non-zero coefficients, and that their coefficient estimates will be normally distributed with the mean estimate being the true coefficient value [42].

2.4 Non-linear models

We next discuss several non-linear methods for regression: random forests, gradient boosting, and support vector machines.

Both random forest and gradient boosting models use *decision trees* to make predictions. A decision tree is a binary tree where each non-leaf node represents a condition and each leaf node represents a prediction value. To make a prediction, start at the root node and check whether the condition at that node is true or false. If true, move down to the node's first child; if false, move to the second child. This process is repeated until a leaf node is reached, which will give the value that the decision tree predicts. Although decision trees can be used as machine learning models on their own, it is more common to use decision trees in *ensemble methods*, which combine many different decision trees into a single model. This is because a single decision tree will usually have high variance; a small change in the training set can lead to a completely different decision tree [21].

A *random forest model* combines many independent decision trees to make one unified prediction [5]. Each tree is fitted independently using a subset of predictors and observations. The predictors chosen for each model are chosen randomly without replacement, while the observations are chosen *with replacement* in a process called *bootstrapping* [13]. To make predictions using a random forest, the predictions for each individual decision tree are first computed. Then, the set of predictions are aggregated to give one final prediction. For regression, one suitable way to aggregate individual predictions is to use the mean.

Boosting is the technique of sequentially improving a weak learner until it becomes a strong learner [31]. Boosting is commonly used with decision trees. Unlike random forest models, where each tree is independent of one another, the trees in a boosting model are fitted sequentially to correct the mistakes made by the previous tree. A *gradient boosting machine* (GBM) is a boosting technique that uses gradient descent to minimize error in a model and correct the shortcomings of previous iterations of the model [17].

The final non-linear model that we considered is the *support vector machine* (SVM) [8, 12]. Support vector machines find a hyperplane that closely fits the data. Data observations that are close to this hyperplane are called *support vectors*, and they have the strongest influence on the model. Unlike linear regression, support vector machines can address non-linear relationships between the response and its predictors.

2.5 Implementation

This section gives the specific details how we fit each model for both the simulated data and the empirical data. Everything in our study was run on version 4.1.0 of `R` [29]. Table 1 summarizes the packages used for each model.

Ordinary least squares models were fitted using the `lm` function from the `stats` package in base `R`. Subset selection models using forward, backward, stepwise forward, and stepwise backward selections were fitted using the `MASS` library. For each of these four algorithms, we fit models using two criteria that determine when to stop adding and removing predictors: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [1, 32]. In general, the AIC will lead to more predictors getting non-zero coefficient estimates.

Table 1: `R` Libraries used and the models used from each library

Library	Models used	Version
`stats` [29]	Ordinary least squares	4.1.0
`MASS` [35]	Forward and backward selection	7.3-54
`glmnet` [15]	Ridge, lasso, elastic-net	4.1-1
`ncvreg` [3]	SCAD and MCP	3.13.0
`xgboost` [7]	Gradient boosting	1.4.1.1
`ranger` [38]	Random forest (simulations)	0.12.1
`randomForest` [24]	Random forest (empirical data)	4.6-14
`e1071` [26]	Support vector machine	1.7-7

Ridge, lasso, and elastic-net models were fitted using `glmnet`. We used the `cv.glmnet` function, which uses cross validation grid search to optimize the penalty scalar λ. Cross validation helps generate a model that performs well on both training and testing data. We used the default value of 10 folds. For elastic-net regression, we used the hyperparameter $\alpha = 0.8$ in our simulation study and $\alpha = 0.5$ in the empirical study. This means that the elastic-net model emphasizes the variable selection provided by lasso in the simulations. The remaining hyperparameters were given their default values.

We used the `cv.ncvreg` function from the `ncvreg` library for SCAD and MCP. Both SCAD and MCP depend on an additional hyperparameter a. We used the default values of a for both models: 3 for MCP and 3.7 for SCAD (the `ncvreg` documentation calls this parameter γ). All other arguments were given their default values.

For gradient boosting and support vector machines, we used cross validation and grid search to find suitable hyperparameters, and then fit a model using the full training set using the hyperparameters selected. Because many of the data sets used had large values of n and p, only a few hyperparameters were tuned. This ensured that the models could be fit within a reasonable amount of time. All other hyperparameters were given their default values.

For gradient boosting with `xgboost`, we varied the learning rate (0.1, 0.3, and 0.5) and maximum tree depth (1, 3, and 7). A maximum of 1000 trees were generated, with an early stopping condition if the model failed to improve for 10 iterations in a row. We used five folds in the cross validation. For support vector machines using `e1071`, we varied ϵ (0.1, 0.5, 2), which affects the model's sensitivity to small errors. We also controlled the cost value C (0.5, 1, 2), which affects how much the model punishes wrong predictions.

With random forests, we used `ranger` for the simulated data. For the empirical data, we instead used `randomForest` because `ranger` could not handle the large number of predictors, resulting in stack overflow errors. For both `ranger` and `randomForest`, we tuned the number of predictors used per decision tree ($\lfloor\sqrt{p}\rfloor$, $\lfloor p/3\rfloor$, and $\lfloor p/2\rfloor$) and the number of trees (300, 400, 500 and 600). The best model was selected based on the out-of-bag error, which represents the average error for each observation using only the trees that did not include that observation.

Some models used could only be used for certain values of n and p. This is because either the runtime becomes infeasible when n or p are large, or the model simply cannot be used when p is too large. Ordinary least squares was only used when $p \leq n$, since it

cannot be used at all when $p > n$. For the same reason, the backward subset selection algorithms were also used only when $p \leq n$. The forward subset selection algorithms were only used when $p \leq n$ and $p \leq 40$. When $p > 40$, the runtimes for forward selection and forward stepwise selection become infeasibly long due to the exponentially increasing number of possible predictor combinations. Lasso, SCAD, MCP, GBM, and random forest models were used for all data sets. Support vector machine models were made for all of the simulated data but was not used for the empirical data because support vector machine models could not handle such a high number of predictors in our empirical data.

3 Monte Carlo Simulations

Monte Carlo simulations use randomly generated data to fit and test regression models. There are several benefits to using simulated data rather than experimental data. For one, the true relationship between the predictor variables and the response is known. Simulations can also be iterated many times, giving sturdier results about the effectiveness of each model. Finally, Monte Carlo simulations give us full control over how our data is distributed. This enables us to evaluate the models under various conditions.

3.1 Simulation Design

Our simulation study used two different functions for the response variable Y. Our first function assumed a linear relationship between the response and its predictors $X_1, X_2, \ldots, X_p$, while the second response used a non-linear relationship. By considering both additive linear and non-linear response functions, we obtain a more thorough understanding of how each model performs in different situations.

The additive linear response function assumes that

$$Y = 1 + 2X_1 - 2X_2 + 0.5X_5 + 3X_6 + \epsilon \tag{7}$$

where ϵ is an independent random error with mean 0 and constant variance. We refer to this linear response function as Model 1. Our additive non-linear response function uses

$$Y = 6 \times 1_{X_1>0} + X_2^2 + 0.5X_6 + 3X_7 + 2 \times 1_{X_8>0} \times 1_{X_9>0} + \epsilon \tag{8}$$

where $1_{X_i>0}$ is the index function given by

$$1_{X_i>0} = \begin{cases} 0, & X_i \leq 0 \\ 1, & X_i > 0 \end{cases} . \tag{9}$$

Note that the non-linear response still includes linear terms. We refer to this non-linear response function as Model 2.

For each simulation, we generated a random $n \times p$ matrix $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ is the multivariate normal distribution with p-dimensional mean vector $\mathbf{0}$ and $p \times p$ covariance matrix $\mathbf{\Sigma}$. The n-dimensional response vector $\mathbf{y}$ was then computed using one of the response functions described in Equations 7 and 8. Finally, a normally distributed random error $\mathbf{e} \sim \mathcal{N}(0, \sigma^2)$ with mean 0 and standard deviation σ was added to each response value.

We assumed that the variance for each predictor was 1, meaning that the covariance matrix $\mathbf{\Sigma}$ is actually a correlation matrix. For every $i \neq j$, the entry $\mathbf{\Sigma}_{ij} \in [0, 1]$ represents the correlation between predictors i and j. The diagonal entries are all equal to 1, indicating that each predictor has variance 1. Correlation between predictors can affect the ability for models to identify important predictors and make accurate predictions.

We considered the following correlation structures for our simulation study:

- *Independent* correlation, where $\mathbf{\Sigma}_{ij} = 0$ for all $i \neq j$;
- *Symmetric compound* correlation, where $\mathbf{\Sigma}_{ij} = \rho \in (0, 1)$ for all $i \neq j$;
- *Autoregressive correlation*, where $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$, where $\rho \in (0, 1)$; and
- *Blockwise correlation*, where $\mathbf{\Sigma}$ is block diagonal with each block having symmetric compound structure (also, each block uses the same value for ρ)

Our simulation study uses a *factorial design*, meaning that we ran simulations using every possible combination of different factors. The factors that we varied in our simulation study are

- The choice of response function (Model 1 or Model 2);
- n, the number of observations (50, 200, and 1000);
- p, the number of predictors (10, 100, and 2000);
- σ, the standard deviation of the random error (1, 3, and 6);
- The correlation matrix structure (independent, symmetric compound, autoregressive, and blockwise); and
- ρ, the correlation between predictors (0.2, 0.5, and 0.9)

By taking every possible combination of these factors, we obtain $2 \times 3 \times 3 \times 3 \times 4 \times 3 = 648$ different settings for the simulations. However, because an independent correlation matrix does not have any correlation between predictors, the value of ρ is not needed. Hence, we only needed to run 540 different settings. For each combination of factors, we ran 100 simulations. Each simulation randomly generated two data sets: one to train the various models, and one to test the models and evaluate performance. Both data sets contained n observations, meaning that a total of $2n$ observations were generated for each simulation.

3.2 Evaluating Model Performance

We used four metrics to evaluate the performance of each model on the simulated data: *train mean squared error*, *test mean squared error*, *β-sensitivity* and *β-specificity*. The mean squared error (MSE) is computed using

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

where y_i is the value of the response and $\hat{y}_i$ is the predicted response value for observation i. In other words, the mean squared error is the average of the squared errors. The mean

squared error was computed on both the n observations used to train the models and the n observations that were not used for training, giving us both a training error and a test error.

Because we are using simulated data, where the true response function is known, we can measure the β-sensitivity and β-specificity for each penalized linear regression model that performs variable selection [25]. A coefficient estimate is a *true positive* (TP) if the coefficient is predicted to be non-zero when that predictor is actually related to the response value. The estimate is a *true negative* (TN) if the coefficient was correctly predicted to be zero when that predictor is not related to the response. A *false positive* (FP) happens when an important coefficient is incorrectly predicted to be non-zero. Finally, a coefficient estimate is a *false negative* (FN) if it was estimated to be zero but that predictor is actually related to the response. A model that perfectly identifies the important and unimportant predictors will have only true positives and true negatives.

The β-sensitivity of a model is given by

$$\beta\text{-sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

while the β-specificity is given by

$$\beta\text{-specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

The β-sensitivity is a measure of a model's ability to correctly identify predictors that are related to the response. If the β-sensitivity is close to 1, then the model assigns non-zero coefficients to all the important predictors; if instead the β-sensitivity is close to 0, then the model cannot identify important predictors well. Similarly, the β-specificity of a model measures how well it can identify unimportant predictors (i.e. predictors that should be given a coefficient of zero).

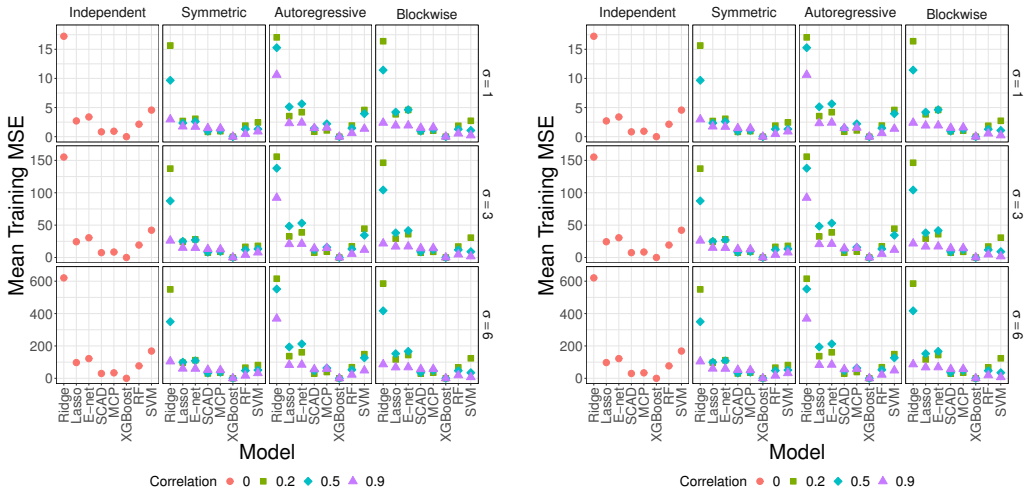
3.3 Linear Simulation Results

Because we ran simulations using 540 different combinations of factors, we only show the results for $n = 50$ and $p = 2000$ in this report (representing the largest ratio between p and n). Results for other combinations of n and p can be found in a supplementary document at github.com/connor-shrader/reu-2021. Each plot measures the average value for one of the four metrics discussed above over 100 simulations. Each row of the plots represent a different value of σ, the standard deviation of the random error. Each column represents a correlation structure. The different shapes and colors for each point represent the strength of the correlation between predictors.

We begin by presenting the results from our simulations for Model 1 (linear function), followed by the results from Model 2 (non-linear function).

Figure 1 shows the average MSE for the simulated models on both training data and test data. Figure 2 displays the β-sensitivity and β-specificity for the linear models that perform variable selection.

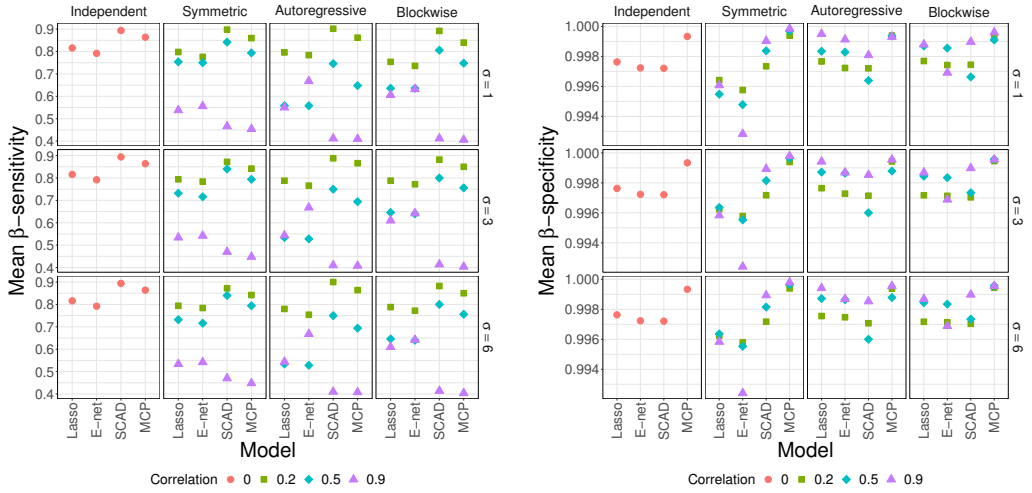
We see that the mean squared error for lasso and elastic-net are generally larger than SCAD and MCP for both the training data and test data. XGBoost has almost zero training



(a) Training data.

(b) Test data.

Figure 1: Average mean squared error for linear simulations when $n = 50$ and $p = 2000$.



(a) β-sensitivity.

(b) β-specificity.

Figure 2: Average β-sensitivity and β-specificity for linear simulations when $n = 50$ and $p = 2000$.

mean squared error under all conditions, but has a relatively large test error. Random forest and support vector machine models have a moderate training error but a large test error. Interestingly, we see that the non-linear models all perform better when there is a strong correlation between predictors. On the other hand, the linear models are somewhat less affected by the correlation.

Looking at Figure 2a, we see that all of the models predict most of the non-zero coefficients when the correlation is low. When the correlation is high, all of the models struggle

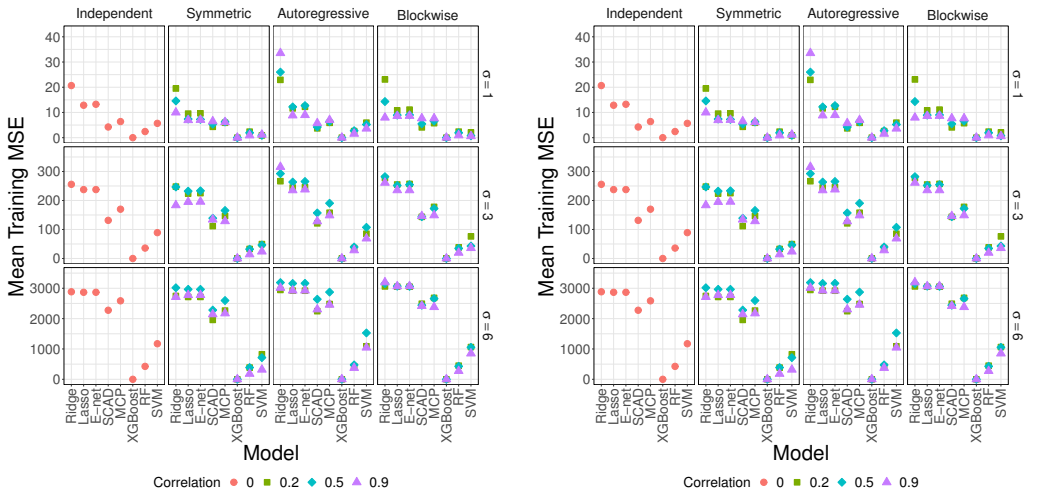
to identify the correct predictors. SCAD and MCP perform the best when the correlation is low but perform the worst when the correlation is high. Elastic-net performs particularly well compared to the other models when the correlation is high, especially when the correlation structure is autoregressive.

Now, consider the results for β-specificity from Figure 2b. MCP appears to make the fewest mistakes when choosing zero coefficients. The performance of the other models depends heavily on the type of correlation and the correlation strength. Lasso and elastic-net perform the worst when the correlation structure is symmetric compound, whereas SCAD performs poorly when the correlation structure is autoregressive or blockwise. We also see that the models correctly identify more coefficients as being zero as the correlation increases.

3.4 Non-linear Simulation Results

Now, we will highlight some results from the simulations for Model 2 given by Equation 8.

Figure 3 shows the average mean square errors on both the training data and test data. We see that the linear and non-linear models have similar test mean squared errors. Another interesting observation is that the non-linear models all have a noticeably lower test mean squared error when the correlation between predictors is high. The linear models still perform significantly worse on the training data compared to the test data.

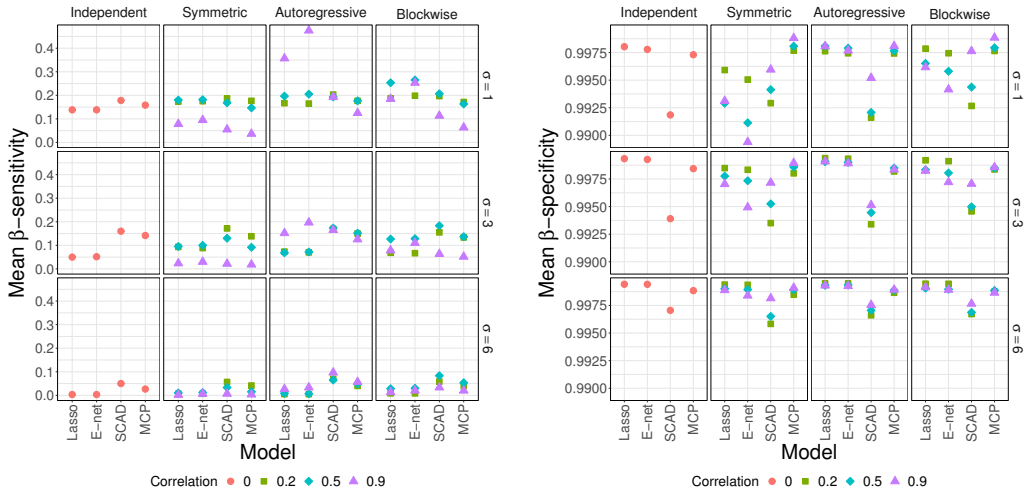


(a) Training data.

(b) Test data.

Figure 3: Average mean squared error for linear simulations when $n = 50$ and $p = 2000$.

Figure 4 shows the results for the β-sensitivity and β-specificity for the non-linear models. We see that all of the linear models estimate almost all of the coefficients as being equal to zero! SCAD and MCP were slightly more likely to correctly estimate non-zero coefficients as being non-zero, but they were also more likely to incorrectly identify unimportant predictors as having non-zero coefficients.



(a) β-sensitivity. (b) β-specificity.

Figure 4: Average β-sensitivity and β-specificity for non-linear simulations when $n = 50$ and $p = 2000$.

4 Empirical Data Analysis

4.1 Details of Empirical Data

For empirical data, we used the Breast Cancer database from The Cancer Genome Atlas (bcTCGA). A cleaned version of the data is provided by the `biglasso` `R` package [39]. This data set contains the gene expression data of 17323 genes from 536 patients. One of these genes is the BRCA1 gene which is among the first genes discovered that can increase the risk of breast cancer [22, 2]. Mutations in BRCA 1 and BRCA 2, another gene discovered 1 year after BRCA1, are responsible for two-thirds of breast cancer cases in women [10]. Because the BRCA1 gene interacts with other genes, it is useful to find genes that interact with BRCA1 to test in further studies [10]. The BRCA1 gene expression level will act as the output value in our regression analysis and the other 17322 genes will serve as predictor values.

This data is a prime example of the large p small n problem where there are many more predictors than data samples. Because of this, only penalized regression and machine learning techniques can be used. This is because there are more predictors than samples which makes least squares linear regression impossible and due to the high number of predictors, subset and stepwise regression becomes too computationally expensive to be feasible. Additionally, support vector machines struggle at such a high number of predictors and resulted in stack overflow errors which made fitting support vector machines on this data impossible. It is also important to note that we do not know whether the response variable is related linearly or non-linearly to the gene expression data. This is why it is important to analyze real, empirical data when comparing mahine learning techniques since we cannot know the functional form of the data.

To evaluate the models, we used *nested cross validation*. We first split the data into five folds. For each of these folds, we used the selected fold as a test set while the other four folds were used as a training set. We then fitted the models using cross validation on this training set, where one interior fold was used as a validation set while the other folds were used to train a model. The role of the validation set in the interior cross validation is different from the test set used in the exterior cross validation. In the interior cross validation, the validation set is used to tune hyperparameters; the model that performs best against the validation set is then chosen. On the other hand, the test set in the exterior cross validation is not used to tune hyperparameters; its only purpose is to evaluate the performance of the models chosen in the inner cross validation. Because the outer test set is not used in the model fitting or selection process, it gives an unbiased evaluation of each model's performance.

We chose to use nested cross validation because it produces five models that were fitted using different subsets of the data for training and testing. If we had only fit one model, the subset of the data we choose for training and testing can have a huge impact on our findings. By using five models that are fit with different subsets of the data, we get a more accurate view of how each model performs in general. Cross validation also allows us to get an idea of how much variance each of these models has by comparing the results between different folds.

The hyperparameters tuned in each of the models were the same as those tuned in the Monte Carlo simulations. For ridge, lasso, elastic-net, SCAD, and MCP, we tuned the penalty strength λ; for elastic-net, we used the hyperparameter $\alpha = 0.5$, meaning that the penalty is in between that of lasso and ridge.

4.2 Empirical Data Results

Recall that we used nested cross validation when fitting models on the bcTCGA data set. This means that we fitted five models using different subsets of the data for training and testing. Figure 5 below shows a plot with the training and test mean squared error for every fold of every model. The bars show the average mean squared error for the five folds. In addition, Table 2 show the aggregated results for the train and test mean squared error.

Table 2: Train MSE, test MSE, and runtime metrics for models fit using the bcTCGA data set.

	Train MSE		**Test MSE**		
Model	**Mean**	**SD**	**Mean**	**SD**	**Mean Runtime** (s)
Ridge	0.1391	0.0266	0.2858	0.0610	29.29
Lasso	0.1842	0.0159	0.2304	0.0337	9.57
E-net	0.1799	0.0127	0.2281	0.0469	**9.41**
SCAD	0.1442	0.0333	0.2218	0.0303	17.28
MCP	0.1566	0.0129	**0.2202**	**0.0224**	15.15
GBM	**0.0002**	**0.0004**	0.2233	0.0507	538.12
RF	0.0378	0.0013	0.2653	0.0525	4906.59

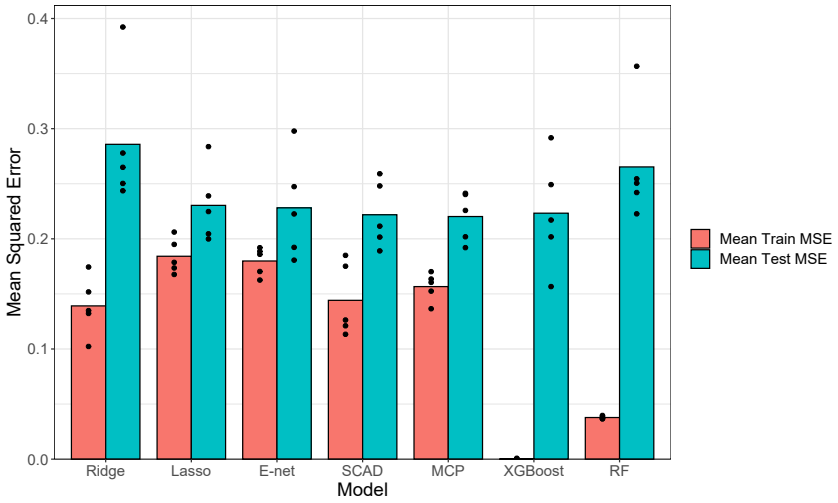


Figure 5: Mean squared error of the models fit on the bcTCGA data set. Each point represents the mean squared error for one fold, while the bars represent the average for the five folds.

5 Discussion

We first describe our key findings from the simulation study. Although results were only shown for the case when $n = 50$ and $p = 2000$, some broad conclusions from other combinations of n and p will be mentioned. We refer the reader to our supplementary document to see the figures and tables for those simulations. Then, we will summarize our findings from the empirical study.

In Model 1 (with a linear response), penalized linear models had a much lower MSE on test data than non-linear models, regardless of the values of n and p. This indicates that linear models are superior to non-linear models when it is appropriate to assume a linear relationship. On the other hand, linear models had a higher MSE on training data than non-linear models in almost all the simulations. We see that the training MSE for the linear models was very close to their test MSE. This indicates that the linear models did not have any overfitting, whereas the non-linear models did. In Model 2 (with a non-linear response), penalized linear models generally have a slightly higher test MSE when $p > n$ when compared to non-linear models. Still, in high-dimensional situations, the linear models are competitive with non-linear models even when the response cannot assumed to depend linearly on the predictors. When $n > p$, the linear models had a significantly higher test MSE. This is expected, since a large number of predictors significantly lowers the variance of the non-linear models. The linear models continue to have a high bias when n is large, which results in large test errors.

For both Model 1 and Model 2, SCAD and MCP generally had the lowest test MSE among the linear models. Subset selection methods that used BIC, where applicable, sometimes performed slightly better than SCAD and MCP. However, the difference was very small. Lasso, elastic-net, and subset selection models with AIC almost never performed as

well as SCAD and MCP, but were usually not significantly worse. Ordinary least squares and ridge regression were the worst-performing linear models. These two models likely suffered from their inability to perform variable selection. Because both Model 1 and Model 2 had only a small number of important predictors, the need for variable selection is pivotal for good model performance. Lasso and elastic-net tended to be more picky about the variables they selected compared to other linear models, especially when the correlation among predictors was high. This is evidenced by these models having a lower β-sensitivity and higher β-specificity. Even though SCAD and MCP were more likely to identify important predictors, all of the linear models struggled to identify important predictors in Model 2. This indicates that when the response cannot assumed to have a linear dependence on the predictors, the linear models may not be reliable for inference.

The standard deviation of the random error did not have any broad qualitative effects on the results, nor did the correlation structure. The standard deviation of the random error did result in worse-performing models, but none of the models appeared to be affected more or less when varying the random error. Overall, we conclude that among the linear models, MCP generally had the best performance. This conclusion backs the results of similar simulation studies that compared penalized regression techniques [4, 40]. When p is large, the performance of MCP is comparable to that of the non-linear models.

Now, we will discuss some findings from the empirical analysis. We found that SCAD and MCP maintained the lowest testing mean squared error among the tested models. This can be seen in Figure 5 and Table 2. XGBoost, elastic-net, and lasso were all had very close performances to SCAD and MCP. Ridge regression and random forest models performed the worst among the models considered. In addition to minimizing the test MSE, MCP also maintained the lowest standard deviation for the test MSE among the five cross-validation folds, as seen in Table 2. Lasso and SCAD also had small standard deviations. On the other hand, the non-linear models had very different performances on each fold, meaning that they are more sensitive to the training data used. These results are all expected, given that the linear models have high bias (resulting in low variance) while the non-linear models have low bias (and consequently high variance). The penalized regression models were fitted exceptionally faster than random forest and XGBoost as documented in Table 2. On average, Lasso and elastic-net ran approximately 56x faster than XGBoost and 510x faster than random forest. MCP and SCAD ran approximately 30x faster than XGBoost and 290x faster than random forest. This provides a significant advantage to the penalized regression techniques, especially given that MCP and other penalized techniques performed better than random forest and XGBoost.

6 Conclusion

There is a severe lack of comprehensive testing comparing traditional machine learning methods such as random forest, gradient boosting, and support vector machines with penalized regression. Our paper bridges the divide between the machine learning and statistical fields in which these two types of models exist in. Testing using Monte Carlo simulations and empirical data has not been tested by other researchers with as many different environments and models.

These comparisons have shown that penalized regression should be added to the tool-

box of any data scientist. In addition to performing better with less error than traditional machine learning techniques, penalized regression is much less computationally expensive. Additionally, in scenarios such as the empirical data study outlined earlier, penalized regression techniques can help determine the relationship between predictors and a response value. In such cases, the ability to determine these relationships can be more important than the predictive performance of a model.

We could also run Monte Carlo simulations where the response is categorical rather than numerical. This could be used to study how penalized regression performs when used for classification data, which is the most common case for high dimensional data sets.

In the future, it may be useful to develop and test a hybrid technique between random forest and penalized regression. This method would harness the power of ensemble learning, while still being able to perform variable selection and would hopefully perform better than either random forest or penalized regression methods individually. Some such models already exist, as seen in [6, 9, 41] which conduct variable selection, however, these models use wrapper methods and variable importance which can be computationally expensive and do not inherently eliminate insignificant variables the same way penalized regression does. Given that random forest models are already very slow, performing additional stepwise selection may result in an exponentially slower runtime. Thus, it is important that inherent variable selection methods such as penalized regression methods are utilized in ensemble methods.

7 Acknowledgments

This research was conducted as a part of the North Carolina A&T State University and Elon University Joint Summer REU Program in Mathematical Biology and was funded by the National Science Foundation DMS# 1851957/1851981.

The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

We would also like to thank Dr. Luke, Dr. Yokley, and Yang Xue for their guidance throughout the research process.

References

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[2] Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, Å., et al. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130, 2003.

[3] Breheny, P. and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.

[4] Breheny, P. and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

[5] Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Capitaine, L., Genuer, R., and Thiébaut, R. Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, 30(1):166–184, 2021.

[7] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. *xgboost: Extreme Gradient Boosting*, 2021. R package version 1.4.1.1.

[8] Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] Degenhardt, F., Seifert, S., and Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 20(2):492–503, 2019.

[10] Deng, C.-X. and Brodie, S. G. Roles of brca1 and its interacting proteins. *Bioessays*, 22(8):728–737, 2000.

[11] Ding, C. and Peng, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[12] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[13] Efron, B. and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

[14] Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[15] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[16] Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[17] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Greene, W. H. *Econometric analysis*. Pearson Education India, 2003.

[19] Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[20] Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[21] James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, volume 112. Springer, 2013.

[22] Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., Jervis, S., Van Leeuwen, F. E., Milne, R. L., Andrieu, N., et al. Risks of breast, ovarian, and contralateral breast cancer for brca1 and brca2 mutation carriers. *Jama*, 317(23):2402–2416, 2017.

[23] Li, Z. and Sillanpää, M. J. Overview of lasso-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and applied genetics*, 125(3):419–435, 2012.

[24] Liaw, A. and Wiener, M. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[25] Liu, X.-Y., Wu, S.-B., Zeng, W.-Q., Yuan, Z.-J., and Xu, H.-B. Logsum+ l 2 penalized logistic regression model for biomarker selection and cancer classification. *Scientific Reports*, 10(1):1–16, 2020.

[26] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2021. R package version 1.7-7.

[27] Miller, A. *Subset selection in regression*. CRC Press, 2002.

[28] Nielsen, D. Tree boosting with xgboost-why does xgboost win "every" machine learning competition? 2016.

[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[30] Sánchez-Maroño, N., Alonso-Betanzos, A., and Tombilla-Sanromán, M. Filter methods for feature selection–a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 178–187. Springer, 2007.

[31] Schapire, R. E. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[32] Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[33] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[34] Usai, M. G., Goddard, M. E., and Hayes, B. J. Lasso with cross-validation for genomic selection. *Genetics research*, 91(6):427–436, 2009.

[35] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[36] Wang, H., Liu, C., and Deng, L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Scientific reports*, 8(1):1–13, 2018.

[37] Wang, L., Chen, G., and Li, H. Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

[38] Wright, M. N. and Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[39] Zeng, Y. and Breheny, P. The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in r. *ArXiv e-prints*, 2017.

[40] Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[41] Zhang, X., Wu, Y., Wang, L., and Li, R. Variable selection for support vector machines in moderately high dimensions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(1):53, 2016.

[42] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[43] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.