Regularization Methods

Kenny Chen *
Department of Mathematics and Statistics, Amherst College

December 13, 2021

Abstract

In the day of big data, the number of predictors often surpasses the number of observations. In these situations, ordinary least squares regression is no longer viable. We explore regularization methods, also known as shrinkage methods, to address the issue of high dimensionality. Shrinkage methods like Ridge, LASSO, and Elastic Network work by "shrinking" coefficient estimates to 0 and sometimes even performing variable selection (LASSO and Elastic Net). We conduct simulation studies to explore how all three perform and then apply it to real world data on milk yield and milk yield related genes. We find that there is no one size fits all solution when it comes to shrinkage methods and that it depends on the makeup of the data.

Keywords: Shrinkage methods, Ridge regression, LASSO, Elastic Net

*The author gratefully acknowledges Amherst College, Dr. Nicholas Horton, STAT495 Peers

1 Introduction

Suppose a biostatistician was interested in examining the linear relationship between gene expression and milk production in cows.

They could consider looking at a linear regression model:

$$Y_i = \beta_{i0} + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i \text{ where } i = 1, \ldots, n \tag{1}$$

which describes the relationship between a response variable Y_i (milk production) and p explanatory variables (genes), $X_{i1}, X_{i2}, ..., X_{ip}$, for $i = 1, \ldots, n$ observations (cows).

This model is near and dear to many of us and for a good portion of the statistics courses at Amherst, we have only learned to fit a linear model using the ordinary least squares method where we minimize the residual sum of squares in order to obtain estimates of $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_p\}$. This fitting procedure has certain advantages, but many alternatives have been proposed to yield better results in two categories:

- Prediction Accuracy: When $n >> p$, that is, the number of observations is greater than the number of predictors, the least squares procedure tends to have low variance but if n is not much larger than p, there can be a lot of variability and if $p > n$, then there does not exist a unique least squares coefficient estimate, meaning the method is not usable. In the day of big data, especially in the field of genomic research, the number of genes as predictors, p, is often much larger than the number of observations, n, at our disposal so this is a prevalent issue. We will examine alternative methods in this paper that address these issues concerning large variance and high dimensionality.

- Model Interpretability: Some methods addressed in this paper will also perform variable selection as it is often the case that some variables in the model are irrelevant. As we will show, some of these methods shrink coefficient estimates to 0 to create a more interpretable model.

We begin by examining three regularization (also known as shrinkage) methods, Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression, and Elastic Net regression that aim to have better results in the above categories by regularizing the coefficient estimates (shrinking the coefficients towards zero). These methods do this by incorporating a shrinkage penalty when minimizing the residual sum of squares. Section 2 briefly explores each method and their advantages and disadvantages. Then simulation results comparing Ridge, LASSO, and Elastic Net are shown in Section 3. Milk yield based genomic data are used to illustrate our methods in Section 4.

2 Methods

All three methods (Ridge, LASSO, and Elastic Net) are based off the ordinary least squares method which minimizes the residual sum of squares defined below but are modified in distinct ways.

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{2}$$

2.1 Ridge Regression

Ridge regression was first proposed by Hoerl & Kennard (1970) in their paper, "Ridge Regression: Biased Estimation for Nonorthogonal Problems". This method is similar to ordinary least squares, but a shrinkage penalty is added on when estimating the coefficients as defined below.

$$RSS + \lambda\sum_{j=1}^{p}\beta_j^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

where $\lambda \geq 0$ is called the *tuning parameter* that is not determined automatically. Breaking down equation 3, we can see that we still want to make the residual sum of squares as small, similar to ordinary least squares, but now we also must consider the shrinkage penalty which has the effect of making $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ shrink towards 0. This shrinkage penalty is based on the ℓ_2 norm of the coefficients ($||\beta||_2 = \sqrt{\sum \beta_j^2}$) which is the distance between the estimates from 0.

The tuning parameter, λ, then tunes the relative influence of these two terms on the estimates. As $\lambda \to \infty$, the coefficients approach zero and when $\lambda = 0$, we are simply minimizing the residual sum of squares. We can also view this process in another formulation as shown below:

$$\underset{\beta}{minimize}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \leq t \text{ for some } t. \tag{4}$$

This states that there exists some t such that 3 and 4 will produce the same $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. When $p = 2$, the Ridge coefficient estimates will have the smallest residual sum of squares out of all the points that are bounded by the circle $\beta_1^2 + \beta_2^2 \leq t$.

2.1.1 Example of Shrinkage

We show a quick example of how the coefficient parameters shrink as λ gets larger. We look at the `mtcars` dataset which contains information on 32 cars from 1973-1974 on their performance and design. We provide Table 1 that is a code book of all the variables.

Suppose we want to predict the cars' miles per gallon using all available predictors using Ridge regression.

In Figure 1, we can see Ridge regression in action. As $\log(\lambda)$ increases, all the coefficients tend towards 0 and no matter how small the coefficient estimates get, they are still nonzero as indicated by the 10's across the top.

Table 1: Codebook for mtcars (n = 32)

Variable	Description
mpg	Miles Per Gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	Engine (0 = V-shaped, 1 = straight)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

2.1.2 Necessity of Standardizing

For Ridge regression coefficients, it is important to standardize the predictors before applying Ridge regression. Consider the previous example where we used various aspects of the car to predict their miles per gallon. The model included the **weight** of the car measured in thousands of pounds. It could very well have been measured in pounds in which case, it would change by a factor of 1000. Lucky for us, most packages in `R` standardize automatically.

2.1.3 Advantages of Ridge Regression

Advantages of Ridge regression can be noticed when we look at it in terms of the bias-variance tradeoff. The bias-variance tradeoff means that Ridge regression will introduce a little bias but in doing so, decrease the variance, ultimately leading to a lower mean squared error and better predictive accuracy.

Figure 2 demonstrates how the mean squared error changes as λ changes for the `mtcars` example from 2.1.1. Mean squared errors (red dots) are displayed with error bars for the lower and upper standard error. Also displayed are two vertical dotted lines. The leftmost line indicates the value of λ that returns the smallest mean squared error. Here, we get the lowest test mean squared error when $\lambda = 2.502$ which we will denote as λ_{min}. The rightmost vertical line indicates the largest λ that is within 1 standard error of λ_{min}, also known as λ_{1se}. People might prefer to use the larger value, λ_{1se}, as a deterrent to overfitting. Moreover, the numbers at the top represent the number of nonzero coefficient estimates.

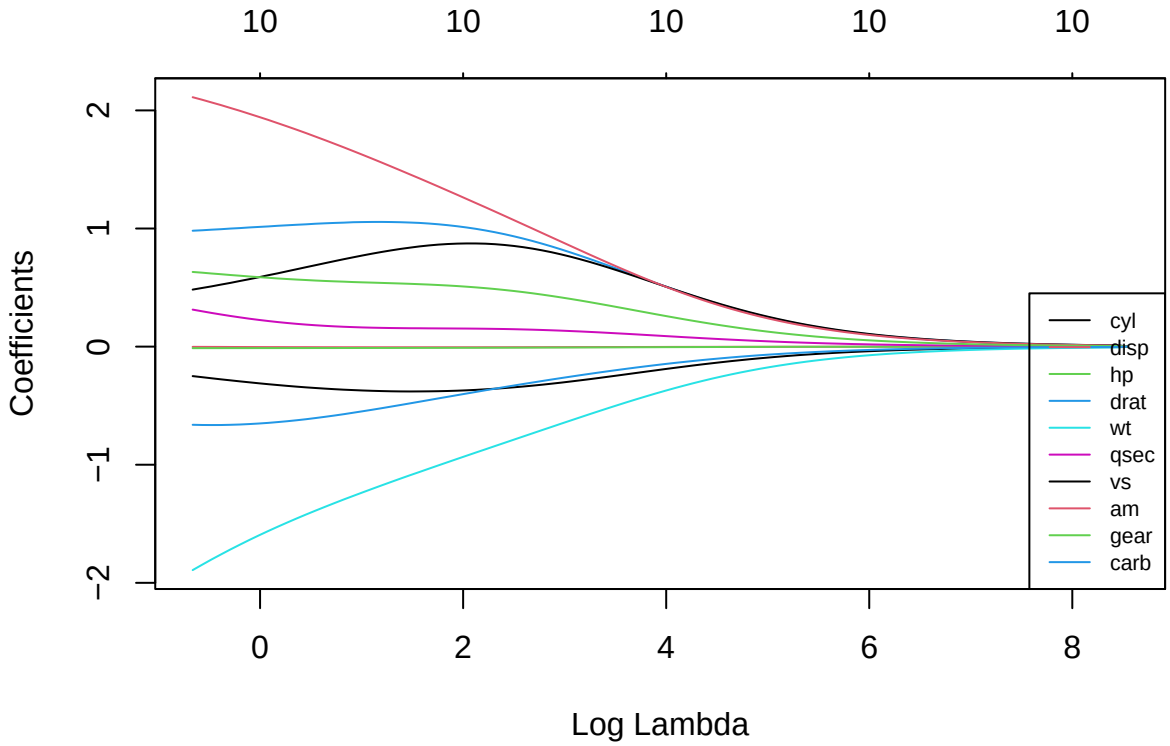

Figure 1: Ridge regression coefficient estimates for `mtcars` as function of Log(λ). The numbers across the top are the number of nonzero coefficient estimates.

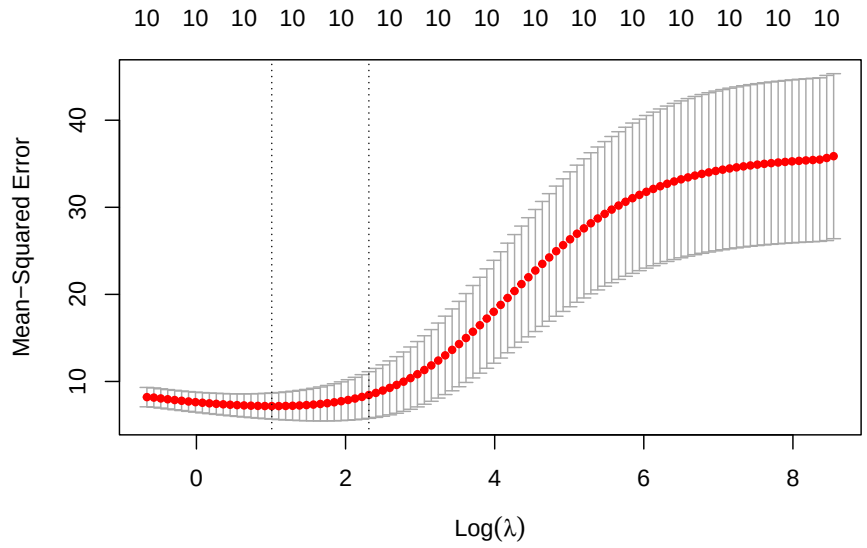

Figure 2: Mean squared errors for the 'mtcars' example in 2.1.1. Mean squared errors (red dots) are displayed with error bars for the lower and upper standard error. Also displayed are two vertical dotted lines. The leftmost vertical dotted line represents λ_{min} which is the value of λ that gives the minimum mean squared error. Rightmost vertical line represents λ_{1se} which is the largest λ that is within 1 standard error of λ_{min}. The numbers at the top represent the number of nonzero coefficient estimates.

In addition to the bias-variance tradeoff, when $p > n$, the least squares estimate will not have a unique solution and the ordinary least squares method will overfit and match the data. Table 2 displays the predicted vs observed `mpg` when we fit an ordinary least squares model on only 7 observations from `mtcars` but still using all variables ($p = 10$). Overfitting is an issue as it makes it difficult for our model to be generalizable.

Table 2: OLS Predicted vs Observed mpg

	Predicted Values	Observed Values
Mazda RX4	21.0	21.0
Mazda RX4 Wag	21.0	21.0
Datsun 710	22.8	22.8
Hornet 4 Drive	21.4	21.4
Hornet Sportabout	18.7	18.7
Valiant	18.1	18.1
Duster 360	14.3	14.3

We can see that the model over fitted and predicted the exact values of the 7 cars. Ridge regression on the other hand regularizes the coefficients to combat against overfitting.

Moreover, Ridge regression can also deal with multicollinearity which is an issue in ordinary least squares. Marquardt & Snee (1975) show how ridge regression is able to estimate coefficients which perform better in predictions than ordinary least squares when multicollinearity is present.

2.1.4 Disadvantages of Ridge Regression

Ridge regression includes all p predictors in the final model. While the shrinkage penalty term will shrink the coefficients towards zero, none of them will be exactly 0 unless $\lambda = \infty$. By including all the predictors in the model, it often leads to uninterpretable models.

2.2 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO regression introduced by Tibshirani (1996), performs variable selection, thus solving the issue that Ridge regression suffers from. The LASSO is very similar to Ridge regression but minimizes the equation below using a different shrinkage penalty:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

In this modified equation, the shrinkage penalty is based on the ℓ_1 norm instead of the ℓ_2 used in Ridge regression. The ℓ_1 norm of a coefficient vector β is given by $||\beta||_1 = \sum_j^p |\beta_j|$.

Moreover, LASSO can also be thought of as solving this problem:

$$\underset{\beta}{minimize}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq t \text{ for some } t. \tag{6}$$

Here we also constrain by some t but instead of the circle defined in 4 when $p=2$, we have a diamond now defined by $|\beta_1|+|\beta_2|\leq t$.

Notice the similarities between Ridge regression and LASSO regression but due to the ℓ_1 penalty, LASSO shrinks the coefficients towards 0 and can even make some be 0 when λ is large enough.

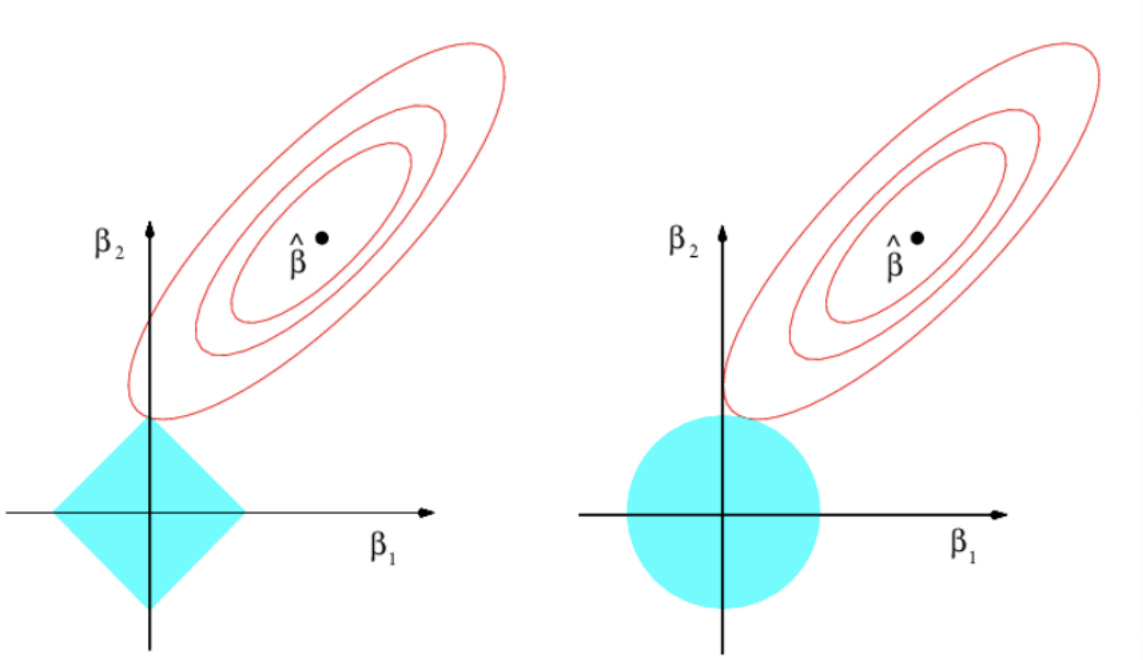

Figure 3: Error and constraint function contours for LASSO (left) and Ridge (right). Shaded blue shapes represent the constraints $|\beta_1|+|\beta_2|\leq t$ and $\beta_1^2+\beta_2^2\leq t$ respectively. Red ellipses represent the contours of the residual sum of squares (James et al. 2021).

Figure 3 demonstrates how LASSO (left) is able to shrink coefficients to 0 while Ridge (right) does not in the 2-dimensional case. $\hat{\beta}$ here marks the ordinary least squares estimates and the red ellipses represent contours of the residual sum of squares. Every point in each ellipse has the same residual sum of squares value and as the ellipses get larger, the value gets larger. Moreover, the blue diamond and circle represent the constraint functions $|\beta_1|+|\beta_2|\leq t$ and $\beta_1^2+\beta_2^2\leq t$ respectively. The point at which the red lines and blue shape touches is the LASSO and Ridge regression coefficient estimates (hence minimizing the residual sum of squares while constraining based off some criteria). Notice that because the LASSO has a diamond shape, the red lines will often touch the diamond on an axis, and when this occurs, one of the coefficients will equal to 0. While James et al. (2021) explains this in the 2-dimensional case, this idea can be extended to higher dimensions.

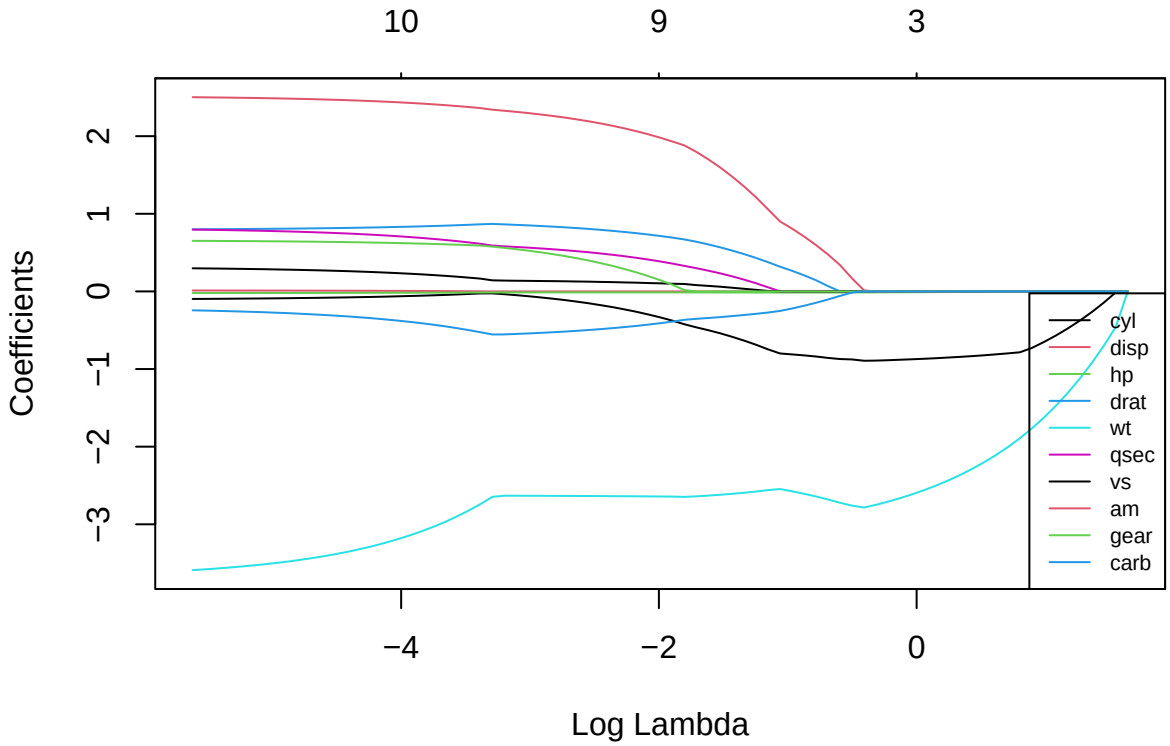

Figure 4: LASSO regression coefficient estimates for `mtcars` as function of Log(λ). The numbers across the top are the number of nonzero coefficient estimates.

2.2.1 Example of Variable Selection

Recall our example of shrinkage in 2.1.1. We can use LASSO instead of Ridge to see how LASSO performs variable selection.

Figure 4 shows how the number of coefficients go from all 10 predictors to 9 then to only three as λ gets increasingly large.

2.2.2 Advantages of LASSO Regression

LASSO's capability to make the coefficients be 0 is a form of variable selection which makes the model sparser and thus more interpretable. On top of variable selection, LASSO has similar advantages that Ridge regression does over ordinary least squares such as shrinkage, generalizability, addressing multicollinearity, and reducing variance.

2.2.3 Disadvantages of LASSO Regression

While LASSO regression solves the issue of model interpretability, it still suffers from some limitations as noted by Zou & Hastie (2005). They consider three scenarios to point out limitations with LASSO regression.

1. When $p > n$, the LASSO selects at most n variables before the model saturates (the model fits the data perfectly due to the number of predictors matching the number of observations). So the model is bounded by the number of observations in our data.

2. Given group of variables with high pairwise correlations, the LASSO tends to select only one variable from the group without caring which one is selected.
3. When $n > p$, if there are correlations between predictors, Ridge regression often performs better than LASSO regression.

2.3 Naive Elastic Net

Elastic Net, the "love child" of Ridge and LASSO regression overcomes many of the issues found in Ridge or LASSO regression and was first proposed by Zou & Hastie (2005). They first introduced the Naive Elastic Net which is defined for any fixed non-negative λ_1, λ_2 as:

$$RSS + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j^2| + \lambda_2 \sum_{j=1}^{p} |\beta_j| \quad (7)$$

Like the others, we can also view it as :

$$\underset{\beta}{minimize} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } (1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} |\beta_j^2| \leq t \text{ for some } t. \quad (8)$$

where we let $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$. Zou & Hastie (2005) calls the function $(1-\alpha) \sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} |\beta_j^2|$ the "elastic net penalty" which is a convex combination of the Ridge and LASSO penalty.

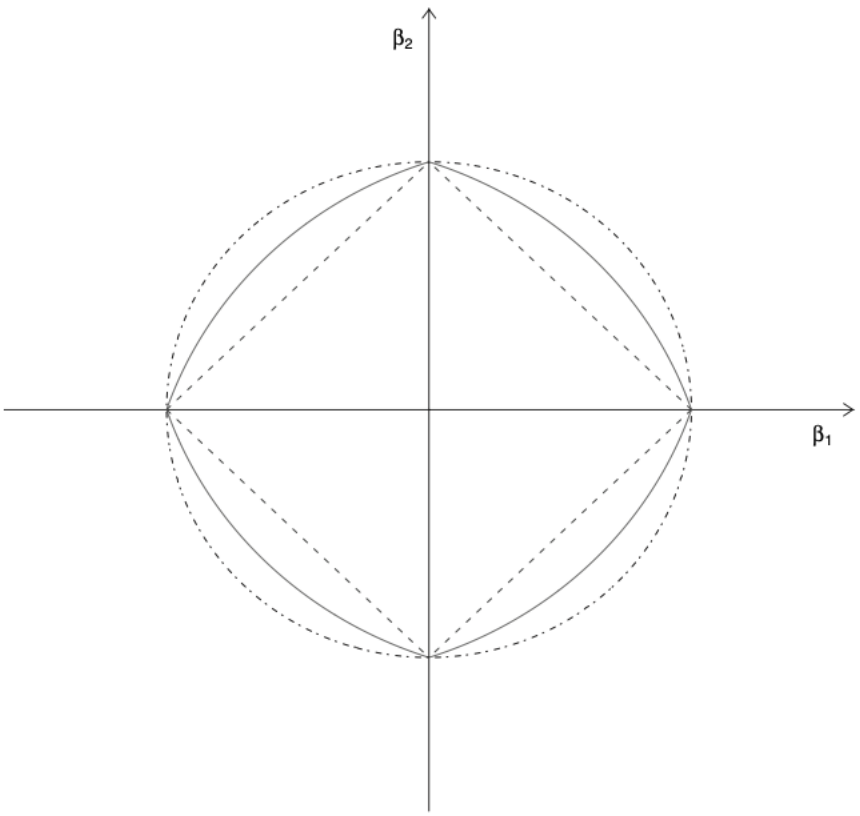

Figure 5: Ridge contour (dotted and dashed circle), LASSO contour (dashed diamond), Elastic Net contour (solid line) (Zou & Hastie 2005).

In Figure 5, we see the constraints of all three methods in 2-dimensions. Notice the naive Elastic Net at $\alpha = 0.5$ is in between the constraints of Ridge and LASSO and depending on α, can become the Ridge ($\alpha = 1$) or the LASSO ($\alpha = 0$) constraint.

2.3.1 Elastic Net

Zou & Hastie (2005) point out that the naive Elastic Net estimator does not perform well unless it is very close to either the Ridge regression or the LASSO regression. Because of this, they propose a rescaled version of the naive Elastic Net:

$$\hat{\beta}(\text{elastic net}) = (1 + \lambda_2)\hat{\beta}(\text{naive elastic net}) \tag{9}$$

2.3.2 Advantages of Elastic Net Regression

In addition to the advantages mentioned above from the other regularization methods, Elastic Net address the issues with LASSO mentioned above. Elastic Net regularization is able to select groups of correlated variables to overcome LASSO's issue of only selecting one variable from a group of highly correlated variables while also shrinking coefficient estimates simultaneously.

2.3.3 Disadvantages of Elastic Net Regression

Due to Elastic Net having two λ rather than just one λ, it becomes more computationally burdensome than LASSO or Ridge. A grid of values for λ_2 are chosen, then for each λ_2, cross-validation techniques are chosen to choose the optimal λ_1 value.

2.4 Selecting the Tuning Parameter

For all three methods, we need to select an appropriate tuning parameter λ. Since it is not done automatically, we can turn to cross-validation to solve this issue. Cross-validation as a technique involves creating equal sized groups where one group is left out at every iteration as the test dataset while the model is fitted on the other groups. This is done so that each group has a chance to be the testing data. We can use this technique here and choose a grid of λ values and find the cross validation mean squared error of each λ value. After cross-validation, we then select the λ that gave us the smallest mean squared error.

3 A Simulation Study

We now perform two simulation studies using the `lassoenet` package. This package utilizes the `glmnet` package created by Friedman et al. (2010) to perform simulations comparing the three methods on simulated data sets. These simulations were performed using `R version 4.1.1` and `RStudio version 1.4.1717`. Here we will show how these models compare in terms of predictive performance using the test mean squared error. To do this, we utilize the `simulation.collinear` function from the `lassoenet` package. This function has 10 parameters that we can alter listed in Table 3.

We will simulate 2 data sets by altering various arguments. We keep `n.resample = 100`, `n = 100`, `matrix.option = 1`, `collinear = 0.5`, `sig = 2`, `split.prop = 0.8`, `step.alpha = 0.2`, `option = 1`, and `parallel = FALSE` constant throughout.

- First Simulation: We simulated 100 datasets with 10 predictors with true coefficients: $\beta = (1, 2, 0, 0, 2, 0, 0, 0, 2, 1)$
  - For this dataset, we wanted to see how the methods would perform when the true coefficient values are only on a select few and there are less predictors than observations.
- Second Simulation: We simulated 100 datasets with 200 predictors where $\beta_i = 0.7$ for all 200 predictors.
  - For this dataset, we wanted to see how the methods would perform when the true coefficient values are constant for all predictors and there are more predictors than observations.

Table 4 displays the mean squared errors for each of the methods in each simulation. In simulation 1 where there a few relevant predictors, the LASSO and Elastic Net were comparable and had lower mean squared errors than the Ridge regression. In simulation

Table 3: Options for simulation.collinear()

Argument	Description
n.resample	The number of simulation datasets to generate
n	number of rows in each dataset
coeff	A vector of true coefficients
matrix.option	1: Use an Exchangeable correlation matrix to simulate the predictors 2: Use an Autoregressive correlation matrix to simulate the predictors
collinear	The correlation levels within the matrix.option
sig	Model variance
split.prop	Specifying training proportion. Testing proportion will be 1 - the training proportion.
step.alpha	The step size of the alpha grid for the Elastic Net
option	1: split the dataset according to c(split.prop, 1 - split.prop) 2: Use the whole dataset. Note: When option = 2, the split.prop will be ignored
parallel	Parallelization

Table 4: Mean squared error of simulations

	Simulation 1	Simulation 2
LASSO	5.379	11.297
Elastic Net	5.315	8.853
Ridge	5.450	5.734

2, we see LASSO had the highest mean squared error with Elastic Net not too far behind, but Ridge performed quite well here when all the predictors were relevant.

4 Real World - Milk Production

We now demonstrate these methods on a real world data set from Seo et al. (2016). They conducted RNA-sequence analysis for milk yield associated genes in cows, in particular Holstein cows. Twenty-one RNA-sequence samples were obtained from the somatic cells of Holsteins' milk and over 13,000 genes were measured as well as milk yield which was measured as "normalized milk yield derived from the Korea Type-Production Index (KTPI)" Seo et al. (2016). More information on the data can be found here. We also refer the interested reader to their paper where they apply different regression methods to their question

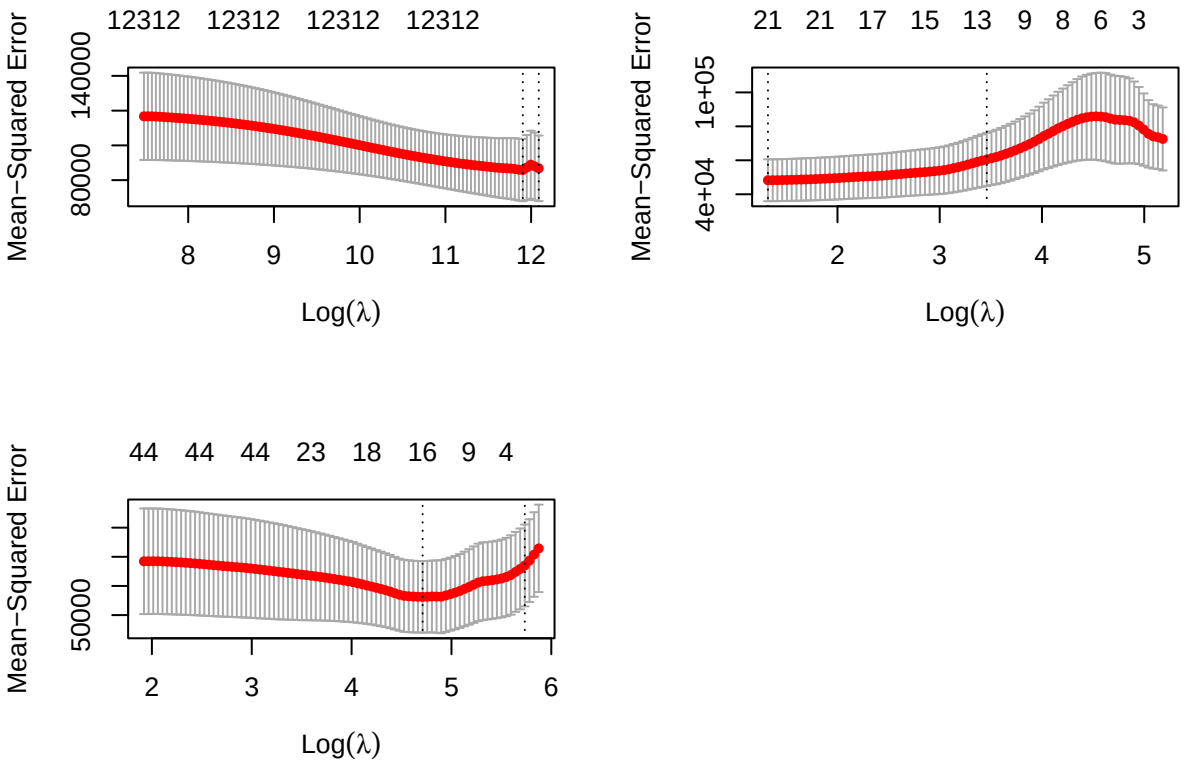

Figure 6: Mean squared errors for Ridge (top left), LASSO (top right), Elastic Net (bottom left).

of interest.

In our case, however, we apply Ridge, LASSO, and Elastic Net regression using all sampled genes, milk parity and lactation period to predict milk yield. Moreover, we also want to see which genes are selected by LASSO and Elastic Net ($\alpha = 0.5$) as important predictors in milk production. We perform 7-fold cross-validation for the 21 observations to select tuning parameters and fit the model. We compare the performance of them based on the cross-validated mean squared error.

Table 5: Mean squared errors from predicting milk production

Method	Lambda	MSE
Ridge	147665.518	85800.81
Lasso	3.744	48252.71
Elastic Net	111.190	56239.43

Looking at Table 5, we see that LASSO performs the best in terms of mean squared error at 48252.71 when $\lambda = 3.74$ and Elastic Net lags just slightly behind while Ridge performs the worst at 85800.81 when $\lambda = 147665.52$. Figure 6 shows how the mean squared error changes as $Log(\lambda)$ changes as well as the number of nonzero coefficient estimates.

Recall that Ridge regression performs better when all predictors influence the response

variable, but because we had over 13,000 genes, LASSO and Elastic Net seems to be more valuable in identifying those that had an impact thus creating a more interpretable and parsimonious model.

Table 6: Selected Variables

LASSO	Elastic Net
BPIFC	BPIFC
CCL14	CACNG2
CYP2E1	FBXO36
DCST1	GALNTL1
FBXO36	GSTA3
GALNTL1	LRRC3
GSTA3	METRN
HS3ST2	MIR2397
IP6K3	PBX1
KLKB1	PCDHGB4
KPNA7	PFN2
METRN	PLIN2
MIR2397	PRMT8
PBX1	PTPRU
PCDHGB4	SARDH
PFN3	SLC4A1
PLIN2	NA
PRMT8	NA
PTPRU	NA
SARDH	NA
STYXL1	NA

Since Ridge regression does not perform variable selection, we turn our attention to the variables selected for LASSO and Elastic Net instead and see how they compare. LASSO regression selected 21 genes while Elastic Net selected 16 genes. Those genes are listed in Table 6.

5 Discussion

We caution the reader when making comparisons between our results and the paper from which this dataset was obtained. We utilized this dataset because of its high dimensionality in the real world which allowed us to show the advantages of these various penalized regression methods as alternatives to the ordinary least squares method which many of us are accustomed to. Moreover, we did not perform any sort of hypothesis testing here. We prioritized predictive accuracy and model interpretability over inference. However,

statisticians have proposed various significance tests specific for these methods such as Lockhart et al. (2014) and Cule et al. (2011) and this as a topic deserves further exploration.

Using simulated data with various characteristics such as high dimensionality, we can see that there is no clear winner when it comes to minimizing the mean squared error. Depending on the data set, the predictive accuracy of each method can change quite drastically.

In general, we can expect LASSO and Elastic Net to perform better when there are a small group of relevant predictors such as in the first simulation while Ridge regression might take the win when there are many relevant predictors and the number of predictors is much larger than the number of observations such as in simulation two. For the milk yield data, we had 21 samples with over 13,000 predictors and we saw that Elastic Net and LASSO performed the best while Ridge had a much larger mean squared error. This could be the case because only few genes were relevant as milk production related genes while many of them were unrelated. Knowing when to use these three methods takes careful consideration of the dimensionality, the relevancy of predictors, multicollinearity and the bias-variance tradeoff. However, knowing that these tools are available as alternative methods to ordinary least squares will certainly prove useful for any statistician.

References

Cule, E., Vineis, P. & Iorio, M. D. (2011), 'Significance testing in ridge regression for genetic data', *BMC Bioinformatics* **12**(1).
URL: *https://doi.org/10.1186/1471-2105-12-372*

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.
URL: *https://www.jstatsoft.org/v33/i01/*

Hoerl, A. E. & Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.
URL: *https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021), *An Introduction to Statistical Learning with Applications in R.*

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014), 'A significance test for the lasso', *The Annals of Statistics* **42**(2), 413 – 468.
URL: *https://doi.org/10.1214/13-AOS1175*

Marquardt, D. W. & Snee, R. D. (1975), 'Ridge regression in practice', *The American Statistician* **29**(1), 3–20.

Seo, M., Kim, K., Yoon, J., Jeong, J. Y., Lee, H.-J., Cho, S. & Kim, H. (2016), 'RNA-seq analysis for detecting quantitative trait-associated genes', *Scientific Reports* **6**, 24375.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.
URL: *http://www.jstor.org/stable/2346178*

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**(2), 301–320.
URL: *http://www.jstor.org/stable/3647580*