Statistical Pattern Analysis Final Paper

Akhil Vasvani

May 4, 2018

Abstract

Stemming from the 1986 Shibata paper, the relationship between sparse estimators and the correct selection of variables as well as the error of the estimation is examined. Already examined in a previous paper, *On the "Degrees of Freedom" of the Lasso*, the author uses either BIC or AIC to select λ, and finds through simulations that BIC performs better to recover the correct support of β than AIC. While this is true, further examination is necessary to see if subsequent inference can be determined. This paper attempts to show that even though BIC will give a better choice to recover the correct support of β, AIC will yield a higher power in test.

1 Introduction

Before diving into sparse estimators, it is prudent to understand Ritei Shibata's paper "Consistency of Model Selection and Parameter estimation." Shibata bridges the relation between the consistency of model selection and that of parameter estimation after a model has been selected. He further demonstrates that if model selection is consistent, then the least order of consistency of the parameter estimate becomes lower than $\sqrt{n}$ [5]. He ultimately proves that model selection is achieved at the cost of a lower order of consistency of the resulting estimate of parameters in some parameter domain [5]. This is an important starting block for sparse estimators because sparsity removes the zero terms from a vector and only the non-zero terms remains—thereby reducing its dimensionality and lowering its order of consistency.

In this paper, the relationship between sparse estimators and the correct selection of variables as well as the error of the estimation is examined. Let's consider the problem of choosing λ in a penalized regression problem in which

$$\hat{\beta}_\lambda = \arg min_\beta(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1)$$

In order to understand the problem, we must first highlight the three different notions of asymptotic convergence: consistency of the coefficient vector $\hat{\beta}$, sparsistency—which in terms of model selection aims to select the correct set of non-zero coefficients, and predictive risk consistency (presistency).

1.1 Consistency

When discussing the quality of estimators in statistics, it is imperative to understand these three common descriptions: consistency, sparsistency, and presistency. Let's review the concepts under the context of linear regression. Suppose that the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$Y_i = \beta^T X_i + \epsilon_i,$$

$Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$. Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ be an estimator of $\beta = (\beta_1, \ldots, \beta_d)$. We define consistency of $\hat{\beta}$ such that

$$||\hat{\beta} - \beta|| \xrightarrow{P} 0$$

as $n \to \infty$.

1.2 Sparsistency

Let's start by defining the support of β to be the location of the non-zero elements:

$$\text{supp}(\beta) = \{j : \ \beta_j \neq 0\}.$$

Then $\hat{\beta}$ is sparsistent if

$$\mathbb{P}(\text{supp}(\hat{\beta}) = \text{supp}(\beta)) \to 1$$

as $n \to \infty$.

1.3 Presistency

Let (X, Y) be a new pair. The predictive risk of β is

$$R(\beta) = \mathbb{E}(Y - X^T\beta)^2.$$

Let $\mathcal{B}_n$ be some set of β's and let β_n^* be the best β in $\mathcal{B}_n$. That is, β_n^* minimizes $R(\beta)$ subject to $\beta \in \mathcal{B}_n$. Then $\hat{\beta}$ is persistent if

$$R(\hat{\beta}) - R(\beta_n^*) \xrightarrow{P} 0.$$

This essentially says that $\hat{\beta}$ predicts nearly as well as the best choice of β.

2 The Lasso

Moreover, the Lasso algorithm fits the three described qualities. But what is the Lasso algorithm?

2.1 The Lasso Estimator

The Lasso estimator—least absolute shrinkage and selection operator—is a commonly used regression analysis method. It performs variable selection and regularization to best interpret and predict statistical models. Moreover, its objective is to solve:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Where t is the free parameter that determines how strict the regularization should be, N is the number of cases, y_i be the outcome and $x_i := (x_1, x_2)$ be the covariate vector up until the i^{th} case. It is commonly found in this form:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where the focus to pick the appropriate λ.

2.2 Choice of λ

However, the issue is the choice of λ has to be different for sparsistency than for consistency and presistency. In *Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso)* [7], Wainwright establishes that a larger λ (let's call it λ_0) is required to achieve correct model selection. Yet, this λ_0 shrinks the non-zero coefficients too much towards zero. Therefore, a smaller λ (call it λ_1) is usually chosen to minimize presistency and achieve consistency.

2.3 Rates of Convergence

In *Sparsity and the Lasso*, Tibshirani discusses the rates of convergence, and shows that $\lambda_0 > c\sigma\sqrt{n \log p}$, while $\lambda_1 = O(\sigma\sqrt{n \log p})$. He also mentions the primal-dual witness method related to the solution of the lasso dual problem. However, it is not a practical algorithm for finding a solution because it requires knowledge of the true support and signs. It instead should be called "primal-subgradient witness method" [6].

2.4 Performing BIC and AIC on Lasso

When performing model selection amongst a set of models, there needs to be a sort of criteria to select the correct one. Hence, Akaike information criteria (AIC) and Bayesian information criteria (BIC) were created. Imagine a set of statistical models for some data. Call k the number of estimated parameters in the model and let $\hat{L}$ be the maximum value of the likelihood function for the model. Then AIC is defined as:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

The preferred model is the one with the lowest AIC value. While AIC rewards goodness of fit—thereby allowing for larger number of parameters, it also includes a penalty that discourages over-fitting. However, for BIC the penalty term is a lot larger than the penalty for AIC as it depends on n, the number of data points, and k.

BIC is formally defined as:

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

In *On the "Degrees of Freedom" of the Lasso*, Tibshirani uses either BIC or AIC to select λ, and finds through simulations that BIC performs better to recover the correct support of β than AIC. However, this is specific to linear regression. Tibshirani states that his algorithm for performing BIC or AIC with the Lasso is used specifically in the context of linear regression. His algorithm is thereby called Adaptive Lasso Shrinkage, which shrinks the data and performs the Lasso algorithm on it. His equation for AIC is:

$$AIC(\hat{\mu}) = \frac{||y - \hat{\mu}||^2}{n} + \frac{2}{n}\hat{df}(\hat{\mu})\sigma^2$$

And his equation for BIC is:

$$BIC(\hat{\mu}) = \frac{||y - \hat{\mu}||^2}{n} + \frac{log(n)}{n}\hat{df}(\hat{\mu})\sigma^2$$

Both are used later in the experiment section. In order to go about analyzing the sparse data (whose coefficients are non-zero), an experiment must be generated to demonstrate the use of BIC versus AIC and the subsequent information. More importantly, what will be shown is in terms of the power of the test: $\hat{\beta}_{\hat{D}_{AIC}} > \hat{\beta}_{\hat{D}_{BIC}}$.

3 The Experiment

Going forward, there are a couple of important declarations that need to be stated. Firstly, the Adaptive Lasso Shrinkage for performing BIC or AIC will be utilized. Secondly, for the experiment, time series regression data is used to build a linear model—in particular credit default rates. This is done because economic data is usually collected by passive observation without the aid of controlled experiments. So modeling the data via linear regression is shown to be extremely useful and will help illuminate the algorithm's effect on modeling sparse data in a linear regression fashion. More importantly, we will use the Adaptive Lasso algorithm to yield the AIC and BIC from the data, which will help select the appropriate λ. After, the power of the test of AIC versus the power of the test of BIC will be determined.

Note Bene: The data was borrowed for the end purpose of the power of the test of AIC versus the power of the test of BIC. There is no fore bound conclusion on the credit default rates.

3.1 Data

Consider a simple multiple linear regression model of credit default rates. The data on investment-grade corporate bond defaults, as well as data on four potential predictors for the years 1984 to 2004 (measured for year t) are as follows [1]:

AGE: Percentage of investment-grade bond issuers first rated 3 years ago.

BBB: Percentage of investment-grade bond issuers with a Standard and Poor's credit rating of BBB, the lowest investment grade.

CPF: One-year-ahead forecast of the change in corporate profits, adjusted for inflation.

SPR: Spread between corporate bond yields and those of comparable government bonds.

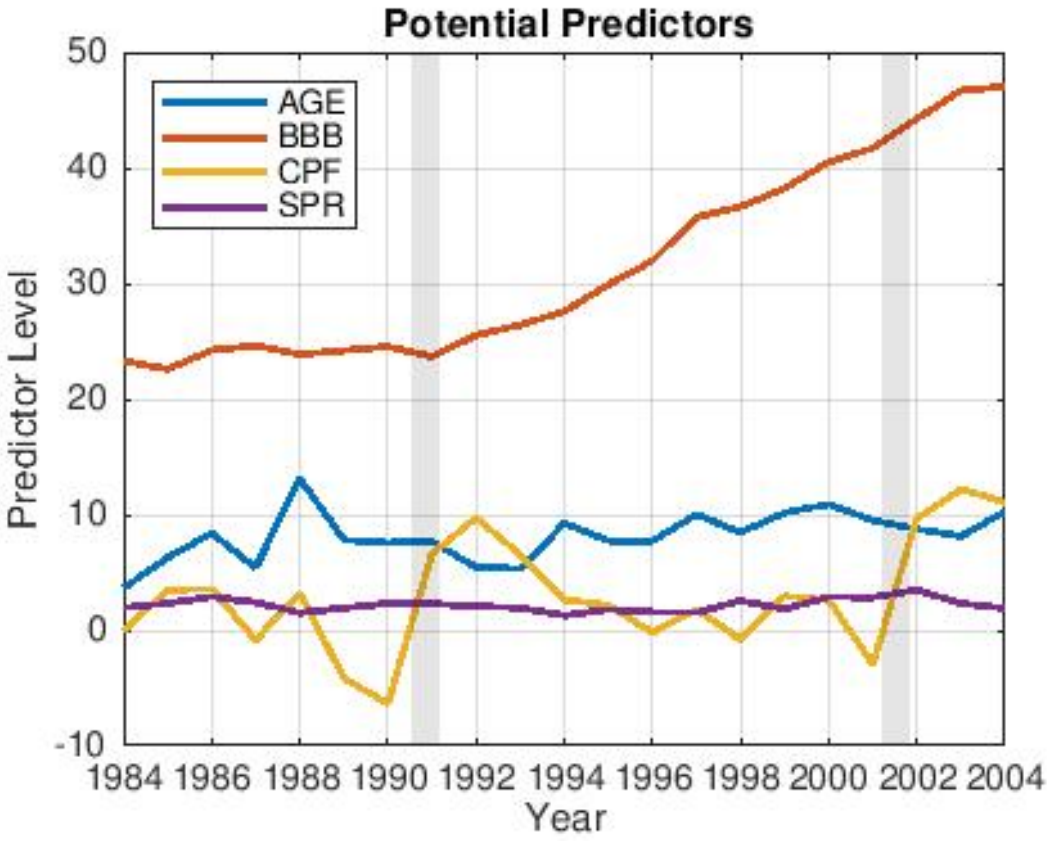

Figure 1: Graph of Predictor level vs Year

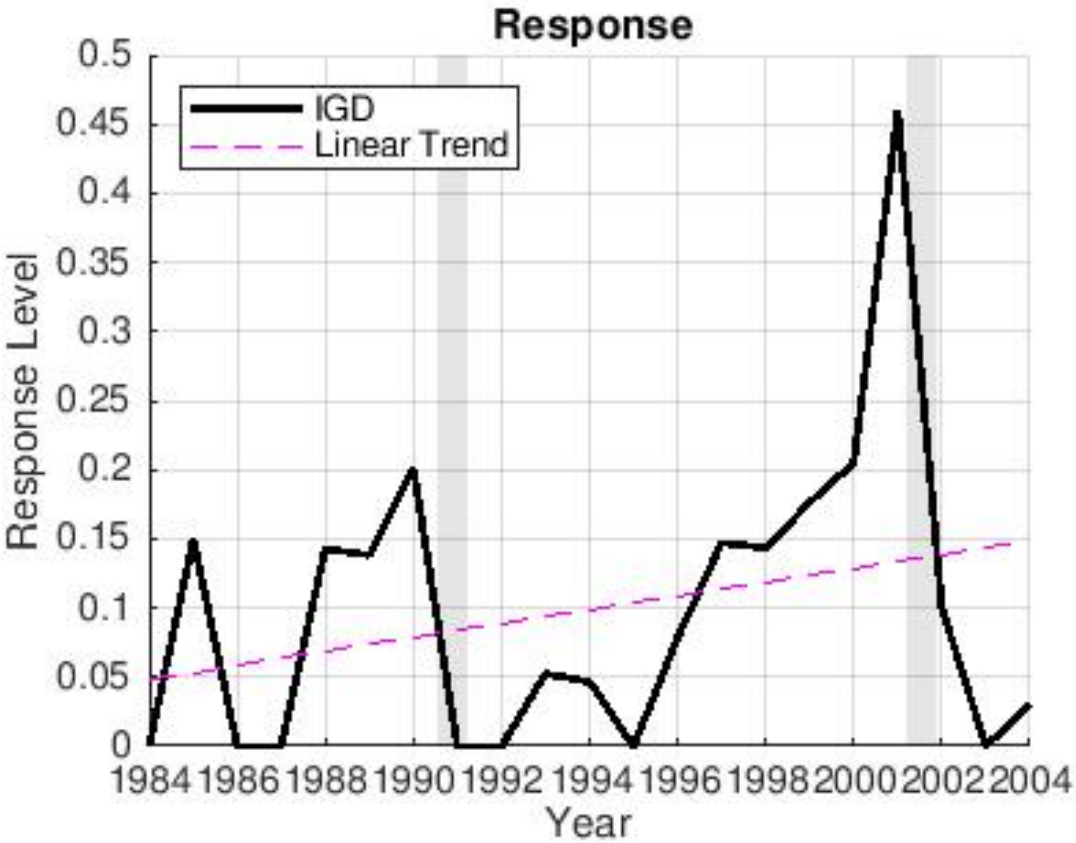

Figure 2: Graph of Response level vs Year

The response, measured for year *t+1*, is **IGD**—default rate on investment-grade corporate bonds.

There are twenty one samples for each predictor. Each predictor was padded with an additional twenty one zeros so the adaptive lasso algorithm could work its magic.

3.2 Methodology

Let's assume the data x is modeled by F_x on $\mathbb{R}^d$. $D \subset [d]$, where d are the signal dimensions on which F_x is non-zero. The model selection estimators are $\hat{D}_{AIC}$ and $\hat{D}_{BIC}$ for AIC and BIC scoring respectively.

Hence, if we assume that $F = X\beta + \epsilon$, then:

$$H_0 : F_x = F$$

$$H_a : F_x \neq F$$

3.3 Results

Before calculating the BIC or AIC scores, a t-test was undertook on the β coefficients to yield several p-values. The β coefficient values were generated from a pre-installed penalized package on MATLAB's main drive [3]. The data was fed in—in this case each predictor's data—and outputted were the β values. Setting the threshold (α) to 0.05, the aim was to determine whether to reject the initial hypothesis or to fail to reject the initial hypothesis.

Let's pick the first three β values for AGE, BBB, CPF, and SPR and show their corresponding p-values.

Table 1: β values for AGE, BBB, CPF, SPR

	AGE	BBB	CPF	SPR
β_1	0.0492	0.0493	0.0493	0.0493
β_2	0.0444	0.0448	0.0498	0.0449
β_3	0.0401	0.0407	0.0502	0.0408

Table 2: p-values values for AGE, BBB, CPF, SPR

	AGE	BBB	CPF	SPR
p-value for β_1	0.4998	0.5000	0.5000	0.4998
p-value for β_2	0.4829	0.4959	0.5046	0.4421
p-value for β_3	0.4641	0.4915	0.5082	0.3797

Clearly, at a significance level of 0.05, all the p-values for each predictor failed to be rejected because they are all greater than α.

Next, the AIC and BIC score were calculated for each predictor. To do this, MATLAB has a pre-installed penalized package on its main drive [3]. The penalized package is an efficient MATLAB toolbox for penalized maximum likelihood. It outputs all the estimator values ($\hat{D}_{AIC}$ and $\hat{D}_{BIC}$) for BIC and AIC scoring as well as their associated λ values in an n x 1 column vectors. Fed in are the four predictors (AGE, BBB, CPF, SPR) and outputted is Table 1.

Table 3: Information Criteria Results for each Predictor

	AGE	BBB	CPF	SPR
AIC	7.6841	7.018	7.8140	7.7069
BIC	4.208	4.2264	4.3387	4.2315

Secondly, amongst each AIC or BIC score is a corresponding λ value. Hence the most optimal λ would have the lowest score. This is represented by Table 2. The package also yielding the optimal λ value.

Thirdly, if not most importantly, the $\hat{\beta}$ values—power of the test—for the estimators $\hat{D}_{AIC}$ and $\hat{D}_{BIC}$ were calculated for each predicator and shown in Table 5. One important fact, these $\hat{\beta}$'s were calculated for fifty samples. They were calculated via a two-tailed t-test comparing each of the estimator's distribution to a value two standard deviations out—both estimators have the same variance. This was done because the estimators themselves are being compared not the actual models. Hence, results were computed and shown in Table 5. Lastly, Table 6 performs the same task as Table 5 except with a larger number of samples.

Table 4: Corresponding λ values for each Predictor based on BIC/ AIC

	AGE	BBB	CPF	SPR
λ	0.002	0.0079	0.0022	0.0002

Table 5: $\hat{\beta}_{\hat{D}_{AIC}}$ and $\hat{\beta}_{\hat{D}_{BIC}}$ for each Predictor for 50 Samples

	AGE	BBB	CPF	SPR
$\hat{\beta}_{\hat{D}_{AIC}}$	0.209 x 10^{-6}	0	0.156 x 10^{-4}	-2.3733 x 10^{-7}
$\hat{\beta}_{\hat{D}_{BIC}}$	0.140 x 10^{-6}	0.2365 x 10^{-6}	0.573 x 10^{-4}	5.7416 x 10^{-8}

Last, but not least, are a couple of the log-likelihoods of the four predictors versus their respective lambda values.

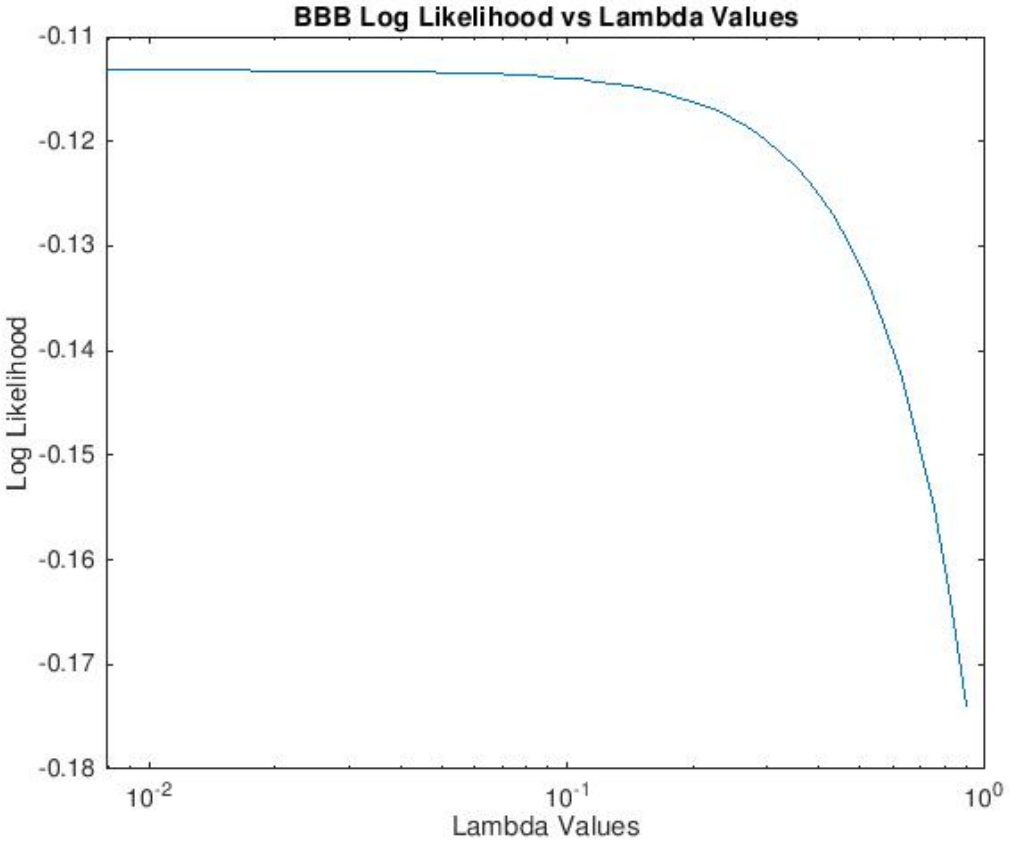

Figure 3: Graph of Log Likelihood vs Lambda Values

Table 6: $\hat{\beta}_{\hat{D}_{AIC}}$ and $\hat{\beta}_{\hat{D}_{BIC}}$ for each Predictor for 2000 Samples

	AGE	BBB	CPF	SPR
$\hat{\beta}_{\hat{D}_{AIC}}$	0	1.75 x 10^{-5}	0	-7.629 x 10^{-6}
$\hat{\beta}_{\hat{D}_{BIC}}$	0.5497 x 10^{-5}	0	0.344 x 10^{-3}	0

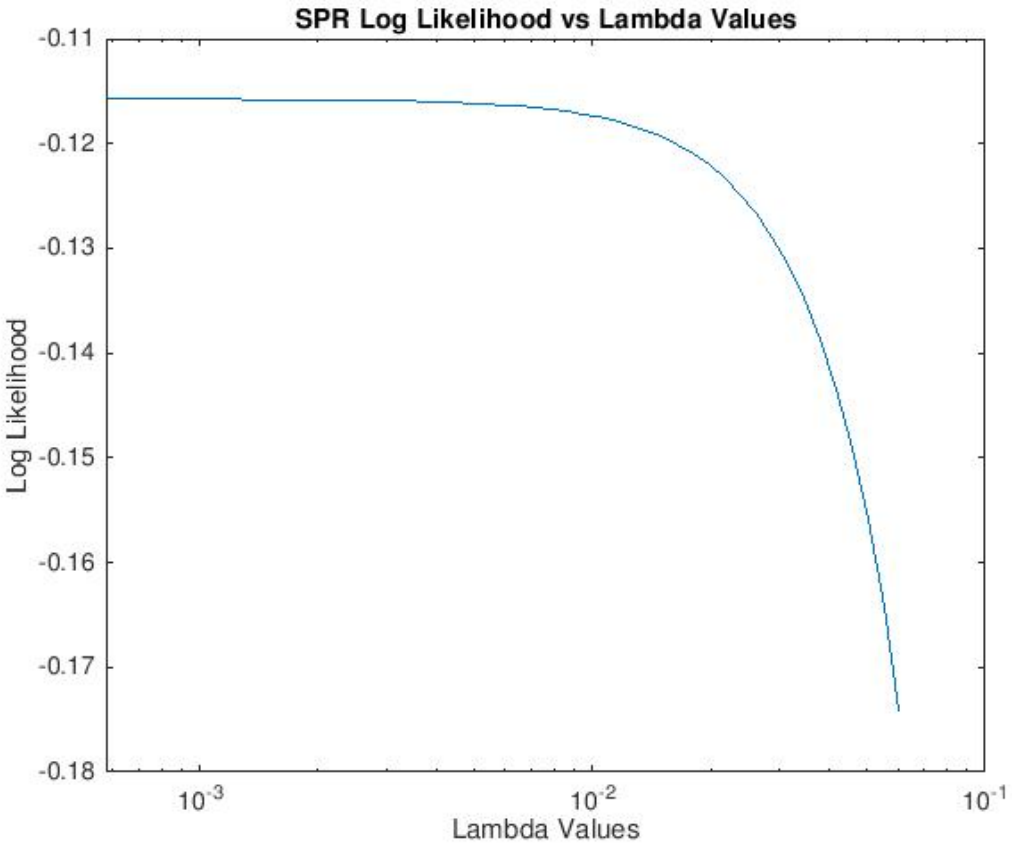

Figure 4: Graph of Log Likelihood vs Lambda Values

4 Discussion and Analysis

To begin the analysis, take notice of Table 3. As per [6], BIC yields a lower score than AIC when recovering the support of β for all the predictors. And accordingly, in Table 4 one can see the optimal λ values are also linked with the lowest BIC and AIC score for each predictor. Table 5 shows the $\hat{\beta}_{\hat{D}_{AIC}}$ and $\hat{\beta}_{\hat{D}_{BIC}}$ for each Predictor for 50 Samples. The first two predictors (AGE and BBB) match the hypothesis: $\beta_{\hat{D}_{AIC}} > \beta_{\hat{D}_{BIC}}$. However, the last two predictors yield contrary results. In fact, one of the AIC scoring for the SPR predictor is negative, which raises the question: what does this mean? At first glance, it could represent that the beta value is stronger than the power. But on a second look, it means that the λ optimal is closer to 0 (as seen in Table 4), so therefore this will throw off the power of the test. Even when observing Table 6 with 2,000 samples, the $\beta_{\hat{D}_{AIC}}$ is still negative. However, unlike the CPR's $\hat{\beta}_{\hat{D}_{AIC}}$ in Table 5, the CPR's $\hat{\beta}_{\hat{D}_{AIC}}$ in Table 6 are greater than the CPR's $\hat{\beta}_{\hat{D}_{BIC}}$. Note that the first two predictors (AGE and BBB) still match the hypothesis.

Included are the two graphs (BBB and SPR) of the log likelihood versus the λ values. Both graphs are monotonically decreasing as expected. However, one slight difference is the range of λ values. In Figure 4, the λ values cutoff before reaching 0.1 while the λ values in Figure 5 cutoff before reaching 1. As seen in Table 4, three of the four predictor λ values are in thousandth values except for SPR—it is in the ten-thousandth. This extremely small λ choice illustrates that small λ yields a low power of the test in Table 5 is almost 0. That said, it does not explain why the

$\hat{\beta}_{\hat{D}_{BIC}} > \hat{\beta}_{\hat{D}_{AIC}}$ for both fifty and two thousand samples. It may be safe to call this result an outlier, but more tests must be conducted to confidently conclude this.

Another aspect that has been examined from "Model Selection" [2] was the mean square error of the each of the predictor's BIC and AIC scores. The initial hypothesis is $MSE(\hat{\mu}_{AIC}) < MSE(\hat{\mu}_{BIC})$ where $\hat{u}$ is the estimator for the mean. This would seem to be counter-intuitive because BIC contains less information than AIC, so hence there would be less variability and would result in lower mean squared error. Using the four predictor's scores, a table was yielded to compare the means, and then MSE was computed.

Table 7: Mean AIC and BIC values for the 4 Predictors

	AGE	BBB	CPF	SPR
μ_{AIC}	7.6996	7.7156	7.8156	7.7203
μ_{BIC}	4.2243	4.2402	4.3403	4.2450

As one can see in Table 7 are the mean values for all four predictors: AGE, BBB, CPF, and SPR. All values are generally around the same—with the lowest BIC value of 4.2243 and the largest BIC value of 4.3403 and similarly for AIC the lowest value of 7.6996 and the largest 7.8156. Now, mean squared error measures the average of the sqaures of errors. Simply put:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x - \hat{x})^2$$

, where x is the measured mean, $\hat{x}$ is the desired mean, and n is the number of samples. Now to prove this result, the MSE will be performed for both BIC and AIC using two different desired values—8 and 6.5 respectively—and show in Table 8.

Table 8: MSE for AIC and BIC

Desired Mean	8	6.5
$MSE(\hat{\mu}_{AIC})$	0.0708	1.5342
$MSE(\hat{\mu}_{BIC})$	13.9714	5.0087

Clearly, the MSE($\hat{\mu}_{BIC}$) > MSE($\hat{\mu}_{AIC}$) from the data provided. $\hat{\mu}_{BIC}$ and $\hat{\mu}_{AIC}$ are the estimators used for the desired mean. However, this is just a base case for the one dimensional case. More simulations are needed for the higher dimensional case to substantiate this.

5 Further Work

In *Relaxed Lasso*, Meinshausen discusses an alternative method is select λ via cross-validation [4]. We will compute this and compare the λ values to the previous λ accordingly:

Table 9: Corresponding λ_{CV} values for each Predictor

	AGE	BBB	CPF	SPR
λ_{CV}	0.0046	0.0383	0.0561	5.7737 x 10^{-4}

As one can compare, Table 7's values are not at all the same as Table 4's. In fact, Table 7's values are all larger. However, while this might obvious, a t-test must be conducted to determine the statistical significance. The β from the Cross-Validation per each predictor must be compared with the β from the previous sections (which match the corresponding λ value in relation to BIC and AIC values in Table 3).

Setting the threshold (α) again to 0.05, it is obvious to see that all the values are greater than the threshold and hence fail to reject that that these are not statistically significant values (see Table 10). Hence, cross-validation does in fact find the minimum λ.

As seen in Figure 5 below, the thin line displays the optimal λ that will minimize the penalty term. However, while this achieves the goal of minimizing the predictive risk, it tends to to include many noise variables in the selected solution [8]. In addition, the accuracy of prediction (in terms of squared error loss) was shown to be negatively affected by the presence of many noise variables,

Table 10: p-values for comparing λ_{CV} against λ

	AGE	BBB	CPF	SPR
p-values	0.6776	0.2886	0.5000	0.5000

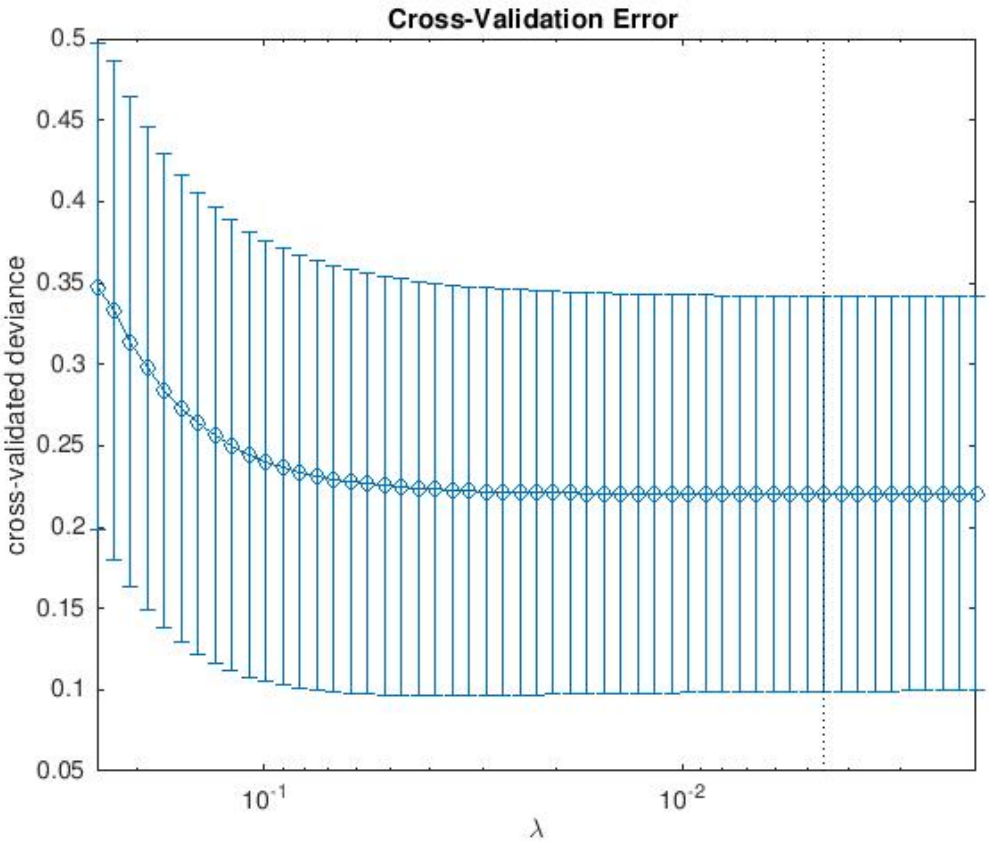

Figure 5: AGE Graph of Cross Validation vs Lambda Values

particularly for high signal-to-noise ratios [8]. Hence, Cross-Validation is a useful technique, but has its own trade-offs as well.

Although it is reasonable to think that the hypothesis has been met as the sample size has increased, there is still more to consider. This experiment conducted was simply a one dimensional case of gathering the power of the test. Another experiment must be conducted in the two dimensional-case. It will be as follows:

Let's assume the data is x is modeled by F_x and the data is y is modeled by F_y all in $\mathbb{R}^d$. $F_x, F_y \in m \subset \bar{M}$ in which M is the model. Since $F_x, F_y \in m$, then there must have the same coefficients: $c_x = c_y$.

Hence then, the idea would be to use BIC to find the estimator $\hat{c} \to c$

$$H_0 : F_x = F_y$$

$$H_a : F_x \neq F_y$$

And ultimately, the goal would be: $\hat{\beta}_{\hat{c}_{AIC}} > \hat{\beta}_{\hat{D}_{BIC}}$. Even though BIC will give the better choice, AIC will yield a higher power of test. This would be a future iteration of this paper should the case arise. A t-test for β would be necessary to yield the p-value which would determine whether to reject or fail to reject the initial hypothesis.

6 Conclusion

In this paper, the relationship between sparse estimators and the correct selection of variables as well as the error of the estimation was examined. For a one-dimensional case, the evidence suggested that though BIC will give a better choice to recover the correct support of β, AIC will yield a higher power in test. To compute this, four linear regression models were all compared. An AIC and BIC was yielded for each. Next, the power of the test ($\hat{\beta}$ values) were calculated. While the data shown implies that the hypothesis is correct, a two dimensional experiment would only provide more insight. In addition, the mean square error of the estimator for the AIC and BIC was compared, which yielded that the mean square error of the BIC was greater than the mean square error of the AIC. This paper's overarching motivation was to start the process to show that better model selection can be achieved with the cost of more bias on the parameter estimates. However, while this result was never achieved in this paper there will be subsequent additions to ultimately go about to prove this goal.

References

[1] Jean Helwege and Paul Kleiman. Understanding aggregate default rates of high yield bonds. 1996.

[2] Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

[3] William H McIlhagga. penalized: A matlab toolbox for fitting generalized linear models with penalties. 2016.

[4] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[5] Shibata, Ritei. Consistency of model selection and parameter estimation. *Journal of Applied Probability*, 23(A):127–141, 1986.

[6] Ryan Tibshirani and Larry Wasserman. Sparsity and the lasso, 2015.

[7] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[8] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.