Penalized Regressions with Different Tuning Parameter Choosing Criteria and the Application in Economics

Sheng Gao *sgao1@samford.edu*
Mathematics and Computer Science Department
Samford University
Birmingham, 35229, USA

Mingwei Sun *msun1@samford.edu*
Mathematics and Computer Science Department
Samford University
Birmingham, 35229, USA

Abstract

Recently a great deal of attention has been paid to modern regression methods such as penalized regressions which perform variable selection and coefficient estimation simultaneously, thereby providing new approaches to analyze complex data of high dimension. The choice of the tuning parameter is vital in penalized regression. In this paper, we studied the effect of different tuning parameter choosing criteria on the performances of some well-known penalization methods including ridge, lasso, and elastic net regressions. Specifically, we investigated the widely used information criteria in regression models such as Bayesian information criterion (BIC), Akaike's information criterion (AIC), and AIC correction (AICc) in various simulation scenarios and a real data example in economic modeling. We found that predictive performance of models selected by different information criteria is heavily dependent on the properties of a data set. It is hard to find a universal best tuning parameter choosing criterion and a best penalty function for all cases. The results in this research provide reference for the choices of different criteria for tuning parameter in penalized regressions for practitioners, which also expands the nascent field of applications of penalized regressions.

Keywords: Penalized Regression, Lasso, Ridge, Elastic Net, AIC, BIC, AICc, Economic Modeling.

1. INTRODUCTION

Regression analysis is widely used to analyze multi-factor data. One of the most commonly used regression methods is linear regression whose estimation can be obtained via ordinary least square (OLS). However, when the dimension of explanatory variables is high, the OLS performs poorly in both prediction and interpretation because of multicollinearity and overfitting effect. In recent years, many admirable penalized regressions have been created and revealed as very useful approaches to fit high-dimensional data because of the ability of performing variable selection and coefficient estimation simultaneously. Therefore, penalized regression methods can find the relationship between the response and explanatory variables and also select out the most significant ones, thereby reducing the dimension of the model. As a result, penalized regression methods can produce models that have stronger predictive performance for the new data because of bias-variance tradeoff (Gunes, 2015). Some popular penalized regression methods include ridge regression (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996), and elastic net regression (Zou and Hastie, 2005). One essential issue of these regularization methods is the choice of tuning parameter which controls the strength of the penalty term. To select the optimal tuning parameter, two commonly used methods include cross-validation (CV) and information criterion such as Akaike's information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978), or AIC correction (AICc) (Sugiura 1978, Hurvich and

Tsai, 1989). However, Chand (2012) showed that the lasso-type methods do not appear to be consistent in variable selection when the tuning parameter is chosen by CV. Similar result can be found in Wang et al. (2009). And the CV approach can also be computationally expensive for big data sets. Thus, it is more interesting to us to investigate the performance of information criteria in penalized regressions. Different information criteria were used in various literatures. For example, Schwarz (1978) has shown that BIC can achieve a suitable trade-off between simplicity and goodness of fit. Shi and Tsai (2002) found that under certain conditions, BIC can consistently identify the true model when the number of parameters and the size of the true model are finite. Ninomiya (2016) obtain the AIC for the Lasso based on its original definition under the framework of generalized linear models. In Fan and Li (2001), both BIC and AIC were applied for tuning parameter selection in their examples. Sen and Shitan (2002) showed the probability of the AICc criterion picking up the correct model was moderately good. Burnham and Anderson (2002) recommended use of AICc as standard compared with BIC and AIC. Due to so many distinct results about these criteria and there does not yet appear to be consensus in literature as to the right approach, in this paper, we explored the performances of various penalized regressions with different tuning parameter choosing criteria in several simulation scenarios and a real data example in economic modeling.

In section 2, we introduced methods of different penalized regressions and tuning parameter choosing criteria. In section 3, simulation examples were shown and different combination of penalized regression and model selection criteria were investigated. Then they were applied in a real economic modeling example. Finally, we concluded in section 5.

2. PENALIZED REGRESSION AND TUNING PARAMETER CHOOSING CRITERION

In this section, we will review some frequently used penalized approaches and tuning parameter choosing criteria for the estimations in existing literature. We consider the usual linear regression model given by

$$y_i = \boldsymbol{x_i^T}\boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where $(\boldsymbol{x_i^T}, y_i)$ is the i-th independently and identically distributed (i.i.d.) random vector, for $i = 1, \dots, n$, such that $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$ is the p-dimensional set of predictor (explanatory) variables, $y_i \in \mathbb{R}$ is the response variable, the $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ is vector of i.i.d. random errors with mean 0 and variance σ^2 and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the vector of parameter coefficients. Let $\boldsymbol{y} = (y_1, \dots, y_n)^T$ and $\mathbf{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$, the model (1) can be written in its matrix form as

$$\boldsymbol{y} = \boldsymbol{X^T}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

The OLS estimates are obtained by minimizing the residual sum of squares. However, OLS often does poorly in both prediction and interpretation because of multicollinearity in data and overfitting effect. Also, the OLS estimates end up with a large variance if the data contains highly correlated explanatory variables (Schreiber-Gregory, 2018). Penalized estimators have been proposed to improve OLS, which minimize the loss function subjected to some penalties.

Hoerl and Kennard (1970) introduced the ridge regression which minimizes the residual sum of squares subject to a penalty on the coefficients with an L_2-norm. The ridge estimator is defined as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{ridge} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X^T}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{\boldsymbol{n}} \left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j \right)^{\mathbf{2}} + \lambda \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \beta_j^2, \end{aligned} \tag{3}$$

where $\|\bullet\|_2$ is the 2-norm of a vector and $\lambda \geq 0$ is the tuning parameter. Ridge regression is promising if there are many predictors which all have non-zero coefficients and are normally

distributed (Friedman et al., 2010). In particular, it performs well with many predictors each having small effect and prevents coefficients of linear regression models with many correlated variables from being poorly determined and exhibiting high variance (Ogutu et al., 2012). However, ridge regression has difficulty in dealing with highly correlated explanatory variables. Moreover, it shrinks the coefficients equally towards zero and never sets them to be exactly equal to zero. Therefore, the ridge regression will not provide the sparse model.

Tibshirani (1996) introduced an L_1-norm penalty and constructed the least absolute shrinkage and selection operator (lasso) estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{lasso} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}\|_{\boldsymbol{2}}^{\boldsymbol{2}} + \lambda\|\boldsymbol{\beta}\|_{\boldsymbol{1}} \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{\boldsymbol{n}} \left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j \right)^{\boldsymbol{2}} + \lambda \sum_{j=1}^{p} |\beta_j|, \end{aligned} \tag{4}$$

where $\|\bullet\|_1$ is the 1-norm of a vector. Unlike ridge regression, the lasso can shrink some coefficients to exactly zero. Hence it does both coefficients estimation and automatic variable selection simultaneously, thereby obtaining a sparse model. However, the lasso is not consistent if there are predictor variables highly correlated. It tends to randomly select one of these variables and ignore the rest.

To overcome the selection bias of lasso estimator, Zou and Hastie (2005) proposed the elastic net estimator by merging the L_1 penalty and L_2 penalty, which is defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{elastic\ net} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta}\|_{\boldsymbol{2}}^{\boldsymbol{2}} + \lambda_1\|\boldsymbol{\beta}\|_{\boldsymbol{1}} + \lambda_2\|\boldsymbol{\beta}\|_{\boldsymbol{2}}^{\boldsymbol{2}} \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{\boldsymbol{n}} \left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j \right)^{\boldsymbol{2}} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{k=1}^{p} \beta_k^2. \end{aligned} \tag{5}$$

The elastic net estimator combines the properties of the ridge estimator and the lasso estimator and is able to simultaneously identify and achieve optimal estimation of the nonzero parameters. Furthermore, the elastic net can select groups of correlated features together when the groups are not known in advance.

Selection of the tuning parameter λ is vital for the performance of aforementioned penalized least squares estimators. It controls the strength of the penalty. Note that linear regression is obtained when $\lambda = 0$. As λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is imposed. When $\lambda = \infty$, all the coefficients are zero. Thus, choosing the optimal tuning parameter is crucial for the penalized regressions to achieve consistent selection and optimal estimation. There are two typically used approaches to select the tuning parameter, i.e., cross-validation (CV) and information-based criterion. As discussed in section 1, we will focus on exploring the effects of latter approach including AIC, BIC, and AICc on the performance of penalized regressions in this paper. BIC is defined as

$$BIC = -2\log\,\mathcal{L}\left(\widehat{\boldsymbol{\beta}}\right) + \log n \cdot K, \tag{6}$$

where $\widehat{\boldsymbol{\beta}}$ is the maximum likelihood estimates of the model parameters, $\log\,\mathcal{L}(\widehat{\boldsymbol{\beta}})$ is the corresponding log-likelihood, n is the sample size and K is the number of parameters of the model. AIC is defined as

$$AIC = -2\log\,\mathcal{L}(\widehat{\boldsymbol{\beta}}) + 2 \cdot K, \tag{7}$$

while AICc is

$$AICc = -2\log\ \mathcal{L}(\widehat{\boldsymbol{\beta}}) + 2\left(\frac{n}{n-K-1}\right)\cdot K, \tag{8}$$

The final terms in (7), (8), and (9) represent a penalty on the log-likelihood as a function of the number of parameters K, which reduce the effects of overfitting. It is not hard to see that the BIC has stronger penalty than AIC for any reasonable sample size n. The formula (9) of AICc can be rewritten as

$$AICc = -2\log\ \mathcal{L}(\widehat{\boldsymbol{\beta}}) + 2\cdot K + \frac{2K(K+1)}{n-K-1} = AIC + \frac{2K(K+1)}{n-K-1}. \tag{9}$$

AICc merely has an additional bias-correction term beyond AIC, which adds stronger penalty than AIC for smaller sample sizes, and even stronger than BIC for very small sample sizes. So, BIC and AICc tend to select smaller models than AIC. Burnham and Anderson (2002) claimed that when n is large with respect to K, then the second-order correction is negligible, and AIC should perform well. And they recommended the use of AICc when the ratio n/K is small (say < 40). In next section, we explore the prediction performance of different penalized methods with the three information criteria in various simulation scenarios.

3. SIMULATION STUDY

This section assesses the effects of different tuning parameter criteria on the prediction performances of aforementioned penalized estimators via various simulation scenarios. All the data sets in this section are generated from the true model

$$y = \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta} + \varepsilon, \qquad \varepsilon \sim N(0,\sigma). \tag{10}$$

$\boldsymbol{x}$ is drawn from a p-dimensional multivariate normal distribution with mean of zero. Within each scenario, the data consists of a training data set and a testing data set. Suppose the size of a training data set is n, then a testing data set of size $n/2$ is generated from the same setting as training data for estimating the prediction performance of the model. To evaluate the performance of penalized estimators, one of the most frequently used measures is the model prediction error (ME) for a model selection procedure which is defined as

$$ME = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{T} E[\boldsymbol{X}^{T}\boldsymbol{X}](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{11}$$

We use the median of the ME (MME) to evaluate the performances of the model selection estimators for a given number of Monte Carlo replications. Specifically, the scenarios which have been investigated are simulated with 50 data sets which is the same number used in Zou and Hastie (2005). The *R* package "*glmnet*" is used to find the penalized estimators and $\alpha = 0.5$ is used for the elastic net in the package. Here are the details of the simulation scenarios.

Scenario 1:

This example was first used in Fan and Peng (2004). Similarly, we let $\boldsymbol{\beta} = \left(\frac{11}{4}, -\frac{23}{6}, \frac{37}{12}, -\frac{13}{9}, \frac{1}{3}, 0, \ldots 0\right)^{T}_{p\times 1}$, and $\sigma = 3$. Hence the number of variables in the true model is $K = 5$. We set the sample size $n = 500$ for each of $p = 5, 20$ and 100. The covariance matrix of the predictor variables is set to $Cov(\mathbf{x_i}, \mathbf{x_j}) = 0.5$ if $i \neq j$ and $Cov(\mathbf{x_i}, \mathbf{x_i}) = 1$, for $i, j = 1, \ldots, p$. In this example, although the dimension of the full model is diverging, the value of n/K is fixed as 100.

The detailed results of scenario 1 is summarized in table 1. As one can see, when all the predictors are correlated and every one of them has effect on the response variable, for example when $p = 5$ in this scenario, the ridge regression performs the best, which is consistent with the

results in Ogutu et al., (2012). Speaking of the ridge regression, one can see that the parameters selected by BIC, AIC, and AICc are the same. This is because the ridge does not produce sparse model. All the predictors are included in the ridge estimators. Thus, the penalties on the number of parameters in the model imposed by these ICs has limited roles. Therefore, they are much likely to select the same estimators. Even for the other two penalized regressions, in many cases the AIC and AICc select the same parameters respectively, see results when $p = 5$ and $p = 20$ of lasso and elastic net. This is consistent with the findings in Burnham and Anderson (2002) that if the ratio n/K is sufficiently large, then AIC and AICc are similar and will strongly tend to select the same model. In this example $\frac{n}{K} = 100$ which is much larger than the threshold (40) given by them. Moreover, from this example we can also see that when there are many redundant variables in the model, the lasso regression with BIC as the tuning parameter choosing criterion outperformed the other penalized methods and criteria, see the results when $p = 100$ which has 95 redundant variables. Also, for a given penalized regression, the BIC has better performance than AIC and AICc when the data contains unimportant variables, for example, when $p = 20$ for lasso and $p = 20$ for elastic net. This is consistent with the property of BIC that it has larger penalty on the dimension of the model than AIC and AICc for large n and can produce a sparser model.

Scenario 2:

In this example, we consider the cases in which $\frac{n}{K}$ also diverges. Similar to Wang et al. (2009), we set $p = \left\lfloor 7n^{\frac{1}{4}} \right\rfloor$ and $K = \left\lfloor \frac{p}{3} \right\rfloor$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. We let $\beta_j \sim U(0.5,1.5)$ for $1 \leq j \leq K$ and $\sigma = 3$. For sample size $n = 100, 400, 1600$, the respective dimensions of the full model and true model are $p = 22, 31, 44$, and $K = 7, 10, 14$. Thus the corresponding floor number of $\frac{n}{K}$ are $\left\lfloor \frac{n}{K} \right\rfloor = 14, 40, 114$. The covariance matrix of the predictor variables is the same as scenario 1.

Method	True model dimension	MME for following IC criteria		
		BIC	AIC	AICc
Ridge	p = 5	8.937	8.937	8.937
	p = 20	9.393	9.393	9.393
	p =100	11.277	11.277	11.277
Lasso	p = 5	9.191	9.191	9.191
	p = 20	9.195	9.213	9.213
	p =100	10.192	10.731	10.695
Elastic net	p = 5	9.139	9.139	9.139
	p = 20	9.297	9.332	9.332
	p =100	10.768	11.199	11.087

TABLE 1: MME values for the simulated scenario 1.

The results of scenario 2 is shown in table 2. Depending on the value of n/K, the performances of AIC and AICc are different. For small n/K, see $n/K = 14$ in the example, AICc outperforms AIC. When n/K is large, see $n/K = 114$ in the example, the results of AIC and AICc are the same. This agrees with the advocate given by Burnham and Anderson (2002) that AICc is preferred to AIC when the ratio n/K is small (say < 40). However, when n/K is sufficiently large, AIC and AICc strongly tend to select the same model. BIC has slight better performance than AIC and

AICc for lasso and elastic net in this example. This is because if the variables have taping effects, which means some β coefficients have real effects but the others taper off quickly to very small values, for example in this scenario there are K coefficients are nonzero and the remains are all zero, then AIC will often choose the very weak effects in a taper. These effects can be estimated very poorly for small sample size. Then zero could be a better estimate for the coefficient, which makes the BIC has a relatively better prediction performance. For the ridge regression, similar results can be seen as from scenario 1 that all IC criteria select out the same parameters.

The following scenarios 3 and 4 were used in the original lasso paper (Tibshirani, 1996) and elastic net paper (Zou and Hastie, 2005), to compare the prediction performance of the lasso and ridge regression systematically.

Scenario 3:
In this example, $\boldsymbol{\beta}$ is specified as $\beta_j = 0.85$ for $j = 1,\cdots,8, n = 20,$ and $\sigma = 3$. The pairwise correlation was set to $corr(\mathbf{x_i}, \mathbf{x_j}) = 0.5^{|i-j|}$. In this example, the sample size is relatively small and the correlation between variable $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ decreases as $|i - j|$ increases.

There are no redundant variables in the true model in this example, which means all the coefficients of the parameters are nonzero. However, unlike the scenario 1 when $p = 5$, the $\frac{n}{K}$ here only equals 2.5 which is much smaller than that of the former (100). The result of this example, presented in table 3, reconfirms our finding in *scenario 2* when $n/K = 14$ that AICc performs better than AIC when n/K is small. And the ridge estimator outperforms the lass and elastic net because all the predictors have effect on the response and the ridge regression does not eliminate any variables.

Method	$\frac{n}{K}$	**MME for following IC criteria**		
		BIC	**AIC**	**AICc**
Ridge	14	10.975	10.975	10.975
	40	9.697	9.697	9.697
	114	9.213	9.213	9.213
Lasso	14	10.638	10.908	10.676
	40	9.596	9.689	9.680
	114	9.210	9.237	9.237
Elastic	14	11.174	11.384	11.277
net	40	9.668	9.712	9.689
	114	9.220	9.249	9.249

TABLE 2: MME values for the simulated scenario 2.

Method	**MME for Following IC criteria**		
	BIC	**AIC**	**AICc**
Ridge	10.928	10.928	10.928
Lasso	12.618	13.241	11.985
Elastic net	11.631	11.997	11.134

TABLE 3: MME values for the simulated scenario 3.

Scenario 4:
This example was created in Zou and Hastie (2005) which contains a grouped variable situation. With $p = 40$, the true model has more redundant variables than example 5. Specifically, we set

$$\boldsymbol{\beta} = (\underbrace{3,\ldots,3}_{15}\ \underbrace{0,\ldots,0}_{25})$$

and σ = 15. The predictors are given by

$$\boldsymbol{x}_i = Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1,\ldots 5,$$

$$\boldsymbol{x}_i = Z_2 + \varepsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6,\ldots 10,$$

$$\boldsymbol{x}_i = Z_3 + \varepsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11,\ldots 15,$$

$$\boldsymbol{x}_i \sim N(0,1), \quad x_i \; is \; i.i.d, \quad i = 16,\ldots 40,$$

where $\varepsilon_i^x \overset{i.i.d}{\sim} N(0,0.01)$, for $i = 1,\ldots 15$. 50 observations are generated in each Monte Carlo repetition. It is easy to see that there are three equally important groups which have five variables within each group and 25 noise features in the model. This example was created to show the superiority of elastic net over lasso when there is group effect among the predictor variables. This can be seen from its result in table 4 that the elastic net outperforms the lasso. We can also see that the AICc has a better performance than AIC and BIC for small sample size and n/K is small ($n/K = 5/3$ in this example).

Method	MME for Following IC criteria		
	BIC	AIC	AICc
Ridge	431.653	431.653	431.653
Lasso	520.568	520.568	512.569
Elastic net	423.631	423.631	420.772

TABLE 4: MME values for the simulated scenario 4.

4. REAL DATA EXAMPLE

In this section, we investigate a real data application and explore the performances of penalized approaches with different IC criteria. We consider the data from Stock and Watson (2005) which contains 540 monthly observations on 131 U.S. macroeconomic time series. Eight of ten major categories of economic indicators are represented within the data. We use the housing starts which is a key economic indicator as an illustration. Specifically, the housing starts of northeast U.S. (HSNE) is set as the response variable. Thus, there are 130 predictor variables. The data is divided into training and testing data sets with the latter has 50 observations. Table 5 reveals the effecting predictor variables selected by lasso and elastic net regressions. We did not list the variables from ridge in the table because it contains all the 130 predictors as the ridge regression does not produce sparse model. The interesting part we found from the result is that all the IC criteria selected the same tuning parameter value in the same penalized regression. Thus, the predictor variables selected by lasso with BIC, AIC, and AICc respectively are the same. So does elastic net. Based on our analysis, we found the reason is the likelihoods produced by lasso and elastic net in this example dominate the penalty terms in the IC criteria formulas. Thus, the penalties imposed on the model size by BIC, AIC and AICc play very limited roles in selection of the tuning parameter. Therefore, they return the same coefficients estimator.

Nevertheless, they provided very meaningful results in this example. For example, as one can see, the variables selected by both lasso and elastic include the housing starts of midwest, south, and west, which means housing starts of the four regions in U.S. are highly correlated to each other. This is consistent with the results in Anaraki (2012) in which the influence of HSNE on housing starts of other regions is described. The Napm Vender Delivery index measures the time for suppliers to deliver essential parts and materials for production. Pring (1992) pointed out that both production index and Napm Vendor Deliveries index have a strong relationship with HSNE and we see both indexes were selected by the lasso and elastic net. Mutikani (2015) showed that an increase in the work force can lead to a growth in house starts. And we see the employment variables in both of the penalized regression. The result also reveals the ability of elastic net to select the grouping variables together. For example, the elastic net includes all the Houses Authorized by Build. Permits of the four regions in U.S.. However, the lasso only contains two of them.

Table 6 displays the prediction mean squared errors of ridge, lasso, and elastic net regression. It can be seen that the lasso selects the least number of predictor variables which is 14, while elastic net has 38 variables in their models. Furthermore, the lasso estimator has the best prediction performance. Its prediction error is as small as 1.995. The ridge regression selects all the 130 predictors and its prediction performance are the worst among them because of overfitting effect.

5. CONCLUSION

In this paper, we investigated the effects of different tuning parameter choosing criteria including BIC, AIC, and AICc, on the prediction performance of some of the most widely used penalized regression approaches such as ridge, lasso, and elastic net, aiming to supplement the existing model selection literature. Both our simulation and real application results support the conclusion in existing literature and provide some guidance to researchers and practitioners who are considering different penalized methods. Using Monte Carlo simulation studies, we compared the prediction performances of the reviewed penalized estimators with different criteria for the choice of tuning parameter. From the simulation results, we find that in general when there are many redundant variables in the model and sample size n is large, then lasso is preferred and BIC is recommended to be used as the tuning parameter choosing criterion for the penalized regression. Because BIC has larger penalty on model dimension than AIC and AICc for large n. It tends to produce a sparser model and can consistently identify the true model for a finite sample size. If grouped effect is found among the predictor variables, then elastic net should be used as it can select all the grouped parameters together while lasso tends to randomly choose one of them. If many predictors have effects on the response variable, one should apply the ridge regression with AIC for large sample size and AICc for small sample size. Because it can keep coefficients of the linear regression model with many correlated featuring lower variance.

We also found that one can use the sample size and the number of nonzero parameters in the model to choose proper criterion for the tuning parameter. When the sample size n or its ratio to the number K of parameters included in the model (i.e. n/K) are small (Burnham and Anderson 2002 used 40 as the threshold for n/K), AICc is recommended for the penalized methods since it has stronger penalty on model size than BIC and AIC under these cases. However, when the ratio n/K is sufficiently large, then AIC and AICc are similar and will strongly tend to select the same model. One shortcoming of using the value of n/K to choose the criterion in practice is that it is hard to know the exact value of K before the regression model is computed. However, it can be used as a supplementary approach for the choice of tuning parameter choosing criterion when the value of K is known. Then our results about it can be used as reference.

We also explored the effect of the tuning parameter choice on variable selection outcomes and prediction results of penalized estimators with a real example in economic modeling. Our result is consistent with existing conclusions found in the study of HSNE. From the real application example,

Economic Indexes Selected	
Both in lasso and elastic net	Elastic net only
1.Industrial. Pro. Index…D.C.G	1.Industial.Pro.index.D.G.M
2.Industrial. Pro. Index…Fuels	2.Industrial.Pro..Index.Res.UTL.
3.NAPM. Pro. Index… Percent	3.Index.Of.W.Adv.In.News.1967.100.SA.
4.Empl.On.Nonfarm.payroll…N.G	4.Emp.Ra.He.W.ADS.No.Unemp.CLF
5.AVE.W.H.of.Pro.Or.Nonsup.W.O.P.NF	5.UNEM.B.Dura.Per.Unemp.27.wks.
6.H.S.Nonfarm.1947...T.F. N1959…Thous.SA	6.AVE.W.Initial.cla.unemp.insur.thous
7.H.S.Midwest.thous.U...S.A.	7.Employ.on.Nonf.Payrolls.Dura.G.
8.H.S.South..Thous.U…S.A.	8.Employ.on.Nonf.Payrools.NonDura.G.
9.H.S. West..Thous.U…S.A.	9.Ave.weekly.hours..mfg.hours.
10.H.Auth.By.Build…Permits,Northeast.Th.U..S.A.	10.H.Author.By,Build.Permit.S.Thou.U.S.A
11.H.Auth.By.Build…Permits.Midwest.Th.U…S.A.	11.H.Author.By.Build..Permits.W.Thou.U..S.A
12.NAPM.Vendor.Delveries.Index.Percent.	12.NAPM.New.Deliveries.Index..percent.
13. Mfrs.new.order.Nondef.capital.g.mil.cha.1982	13.NAPM.Inventories.Index..percent
14.fygm3.fyff	14.Com.Industr.Loans.Oustanding.In.1996.Dol
	15.WKLY.RP.LG.Com.L.Ban.Net.Cha..Indus.L.B
	16.S.P.S.Compo.Comm.Stock..Divid.Yie.Per.AN
	17.Inte. Rate..U.S.Trea.Bill.SEC.M.3.Mo..P.A.N
	18.Inte.Rate..U.s.Trea.Con.Matur.1.YR..P.A.N
	19.fygt1.fyff
	20.fybaac.fyff
	21.United.States.Eff.Exchan.Rate.M..Index.NO..
	22.Foreign.Exchange.rate.JAPAN..Yen..Per.U.S.
	23.Foreign.Exchange.rate.Unite.Kingdon..C.P.P
	24.U..Of.Mich.Index.Of.Consumer.Expecta.BCD.83

TABLE 5: Variables selected by lasso and elastic net with BIC, AIC and AICc.

Method	Number of selected Variables	Prediction Error of Test Data
Ridge	130	362.156
Lasso	14	1.995
Elastic Net	38	2.408

TABLE 6: Prediction Error of Penalized Regression.

we can see that problems from real practice can be very complex. In practice, we are not likely to make the collected data exactly satisfy the data structures in simulation cases. Based on our results, we suggest that if the practitioner intends to get a sparse model with good prediction performance, the lasso estimator is preferred. If the practitioner does not want to ignore any variables in the same group, then elastic net should be used. BIC can be used as the tuning

parameter choosing criterion in the penalized regression in order to obtain a sparser model for a large sample size and AICc for a small sample size. Through this paper, we focused on data with univariate response. To investigate the effects of different tuning parameter choosing criteria on various regularization methods when the data contains multivariate response variables will be an important area for our future work. Moreover, we will expand this research to more regression models, for example, logistic regression and Poisson regression models. These future researches can have significant meanings in theoretical analysis and real applications in this area.

6. ACKNOWLEDGMENTS

This research is supported by Faculty Development Grant FDG085 from Samford University.

7. REFERENCES

[1] H. Akaikei. “Information theory and an extension of maximum likelihood principle,” in Proc. 2nd Int. Symp. on Information Theory, 1973, pp. 267-281.

[2] N. Anaraki. “A housing market without fannie mae and freddie mac: The effect of housing starts.” The Heritage Foundation. Oct. 2012.

[3] K. Burnham and D. Anderson. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. NY: Springer, 2002.

[4] S. Chand. “On tuning parameter selection of lasso-type methods-a monte carlo study.” In Proc. IBCAST, 2012, pp. 120-129.

[5] J. Fan and R. Li. “Variable selection via nonconcave penalized likelihood and its oracle properties.” Journal of American Statistical Association, vol.96, pp. 1348-1360, Dec. 2001.

[6] J.Fan and H. Peng. “Nonconcave penalized likelihood with a diverging number of parameters.” The Annals of Statistics, vol.32, pp.928-961, 2004.

[7] J. Friedman, T. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent.” Journal of statistical software, vol.33, pp. 1-22. Aug. 2010.

[8] F. Gunes. “Penalized Regression Methods for Linear Models in SAS/STAT®.” in Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS Institute Inc. [on-line] Available: http://support. sas. com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels. pdf. 2015.

[9] A.E. Hoerl and R.W. Kennard. “Ridge regression: Biased estimation for nonorthogonal problems.” Technometrics, vol.12, pp. 55-67, Feb. 1970.

[10] C.M. Hurvich and C.L. Tsai. “Regression and time series model selection in small samples.” Biometrika, vol.76, pp. 297-307, Jun. 1989.

[11] L. Mutikani. “Housing starts near eight-year high, but permits fall.” Internet: https://finance.yahoo.com/news/uhousing-starts-near-eight-132259302.html, Aug.15, 2015 [Mar.26, 2020].

[12] Y. Ninomiya and S. Kawano. “AIC for the LASSO in generalized linear models.” Electronic Journal of Statistics, vol. 10, pp. 22537-2560, 2016.

[13] J.O. Ogutu, T. Schulz-Streeck, and H.P. Piepho. “Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions.” BMC proceedings, vol.6, pp. S10, Dec. 2012.

[14] M.J. Pring, The all-season investor: successful strategies for every stage in the business cycle. John Wiley & Sons, 1992.

[15] D.N. Schreiber-Gregory. “Ridge Regression and multicollinearity: An in-depth review.” Model Assisted Statistics and Applications, vol.13, pp. 359-365, Jan. 2018.

[16] G. Schwarz. “Estimating the dimension of a model.” The annals of statistics, Vol. 6, pp.461-464. 1978

[17] L.K. Sen and M. Shitan. “The performance of AICC as an order selection criterion in ARMA time series models.” Pertanika Journal of Science and Technology, vol.10, pp.25-33. Jan. 2002.

[18] P. Shi and C.L. Tsai. “Regression model selection—A residual likelihood approach” Journal of the Royal Statistical Society Series B, vol.64, pp. 237-52, May. 2002.

[19] N. Sugiura. “Further analysts of the data by akaike’s information criterion and the finite corrections: Further analysts of the data by akaike’s.” Communications in Statistics-Theory and Methods, vol.7, pp. 13-26, Jan. 1978.

[20] R. Tibshirani. “Regression Shrinkage and Selection via the Lasso.” Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, pp. 267-88, Jan.1996.

[21] H. Wang, B. Li, and C. Leng. “Shrinkage tuning parameter selection with a diverging number of parameters.” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 71, pp.671-683, Jun. 2009.

[22] H. Zou and T. Hastie. “Regularization and variable selection via the Elastic Net.” Journal of the Royal Statistical Society: Series B, vol.67, pp.301-320, Apr. 2005.