Predicting Median House Value

In this mini project, we are considering 4 different methods on the Boston Housing data found at `http://archive.ics.uci.edu/ml/datasets/Housing`. The shape of the data is 506 observations by 14 attributes. The full text describing the sources, past usage, relevant information and Attribute information can be found `http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names`. Also, all R code for this project is at `https://github.com/ljstrnadiii/Miniproj3`

There are 506 observations with 14 attributes.

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $\$10,000$
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's$

We will consider 4 models including: Principal Component Regression, Lasso Regression, Ridge Regression, and Partial Least Squares Regression. We do not do any prior predictor subset selection processes. We standardize the predictors first before using each model. The focus here is to use four different methods of model selection in order to choose the regularization parameters, λ and t, along with number of components for PLS and PCR. The four methods are the normal training and test procedure, AIC, BIC, and Cross Validation.

We fit each of the four models and get the corresponding mean prediction error for the testing and training procedure, AIC, BIC, and CV estimate for expected prediction error. Briefly reviewing the table summary of the result we see that CV and AIC seem to yield consistent parameters. We know that CV also yields the most unbiased estimate of the actual prediction error. Lasso seems to perform best between AIC and CV. At first glance, I would choose the Lasso with AIC.

Table 1: Parameters with error over minimum mse, AIC, BIC, CV.

Method	test and train	AIC	BIC	CV
Ridge	24.70 , $\lambda = 18$	24.85, $\lambda = 4.0$	557.83, $\lambda = 300$	23.05, $\lambda = 4.0$
Lasso	25.79 , $t = .34$	22.98, $\lambda = .99$	515.45, $t = .01$	23.16, $t = .99$
PCR	24.96 , $p = 4$	23.05, p=13	510.23, p=1	23.5, k=5, p=13
PLSR	25.33 , $p = 2$	23.05, p=13	510.23, p=1	24.0, k=5, p=13

Table 2: Plot-suggested parsimonious model selection.

Method	test and train	AIC	BIC	CV
Ridge	*	*	*	*
Lasso	*	< 28, t=.40	*	<28, t=.40
PCR	*	*	*	<27, k=5, p=11
PLSR	*	*	*	<28, k=5, p=9

The test and train method is the simplest method of approximating the expected predicted error. The benefits of using this method is that there is no computational cost of resampling the data and fitting the model as many times as the other methods presented here. The downside of using this method is that we are using less information from the data to train the model. Holding out a test set and training the model on a subset of the data has the most significant difference with PCR and PLSR. There are significant discrepancies between AIC and CV which more or less get to use more information to fit the model.

The AIC method is a great method to use when we do not have extra data to create an independent testing data set to approximate the expected prediction error. It agrees with CV in terms of the error and parameter tuning and is computationally more simple than CV to compute. The plots of AIC and Ridge, PCR, and PLSR are rather steep and in order to reduce complexity we would have to sacrifice significant error. However, the AIC and Lasso plot seems to suggest that we can reduce t without much of a sacrifice in error. The second table above suggests a value of t that would lead to a more parsimonious model. Since AIC agrees with the more unbiased estimating approach, CV, it is a contender in being a good parameter tuning approach.

The BIC method does not perform well on any of the models. BIC is strongly decreasing towards greatest λ for Ridge and smallest t for Lasso. This means that the models with minimum BIC are the models that shrink the coefficients to 0. The models are being over penalized. The downside to BIC is that it is less tolerant than AIC as N increases, specifically when $ln(N) > 2$. We see that PLR and PLSR are being over-penalized as well as BIC only suggest to use 1 component. BIC does best when the proportion of parameters to N is large. We should refrain from using BIC in this context.

The CV method estimates the expected prediction error while being able to use more data to fit the model. Choosing the appropriate number of folds is another consideration which makes this a bit more involved. Additionally, CV is computationally expensive as it is required to resample the data and fit a model k times. Referring to the CV plots below we see similar plots to AIC for Ridge and Lasso. The plot of Lasso with CV suggests a similar parsimonious model that the AIC plot suggests. We look at various k-fold approaches for PCR and PLSR and do not see much of an increase in variance in the plots. Regardless, the table above considers k=5 folds in the CV process. The plot-suggested table provides the more parsimonious model by considering fewer number of components while retaining comparable CV-error. Choosing the CV method for PCR and PLSR seems like it could have great potential in some contexts. In some contexts, reducing the number of components is key in reducing the complexity of the problem.

If restricted to the test and train method it would be best to choose Lasso because it agrees the most with AIC and CV. The parameter tuning is not as significantly different between tuning methods. If restricted to AIC, it would be wise to choose Lasso, again, for the same reason. If restricted to BIC, PCR or PLSR would be best. If restricted to CV, Ridge would be best because it has more of a shrinking effect than Lasso, $\lambda = 4$ vs $t = .99$, on the coefficients which enforces smoother fitted hyperplanes. This reduces the chance of overfitting. It also has the minimum CV-error.

Next, we discuss which tuning method is best among all models. AIC and CV seem to be fairly consistent. the benefit of CV is that is attempts to estimate the expected prediction error and tends to be less bias than AIC. The downside of CV is that it is a bit more computationally involved. The CV-error plot seems to suggest parsimonious models more so than AIC. Even though AIC can be a bit more bias it is easier to compute. Depending on the context the advantage of CV expected prediction error estimation may be less of an advantage than AIC's quick calculations. In conclusion, I would choose to use AIC and CV. I would use AIC to tune parameters and perhaps use CV to find its estimated expected prediction error.

Plots

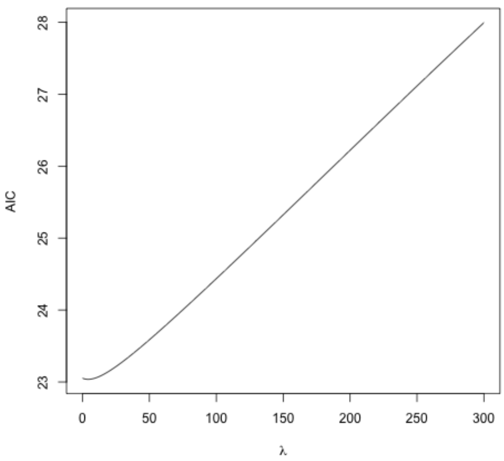

Figure 1: AIC: ridge

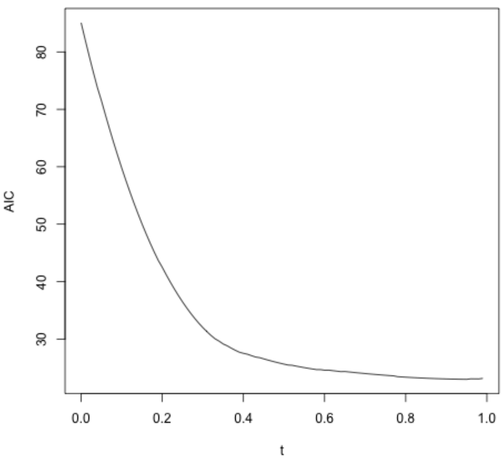

Figure 2: AIC: Lasso

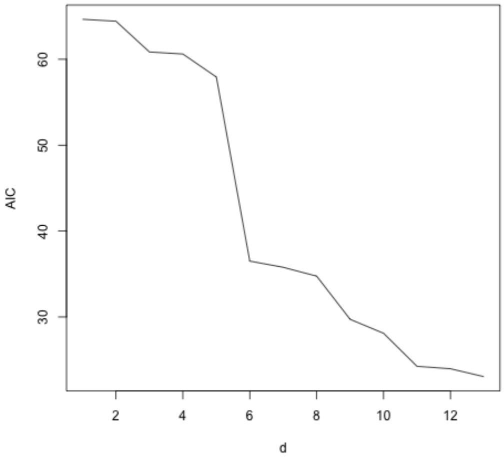

Figure 3: AIC: pcr

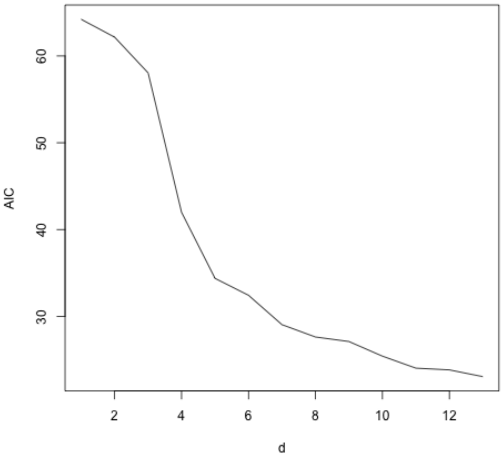

Figure 4: AIC: pls

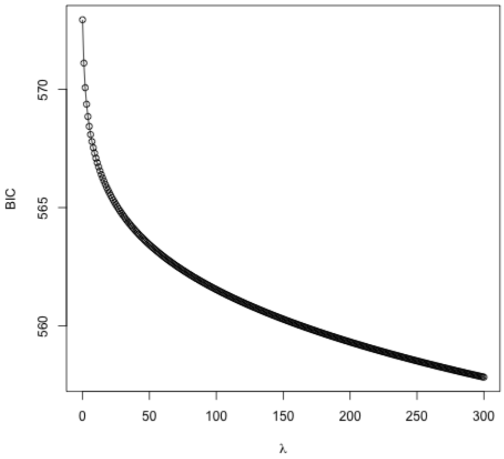

Figure 5: BIC: ridge

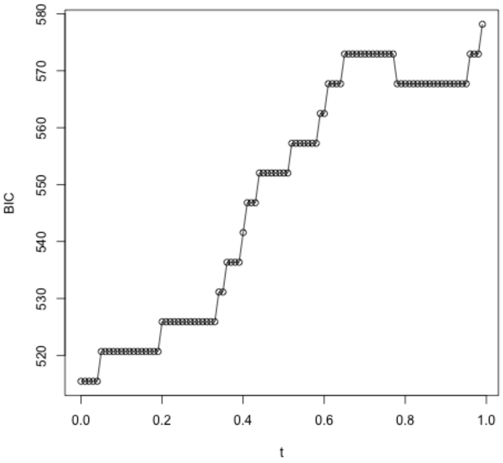

Figure 6: BIC: Lasso

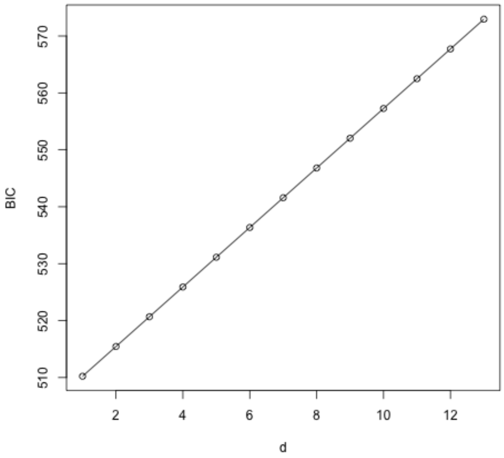

Figure 7: BIC: pcr

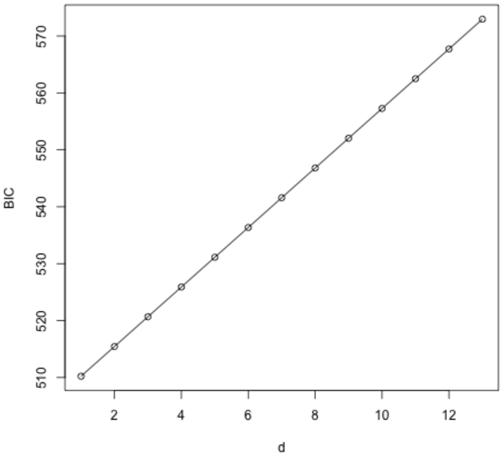

Figure 8: BIC: pls

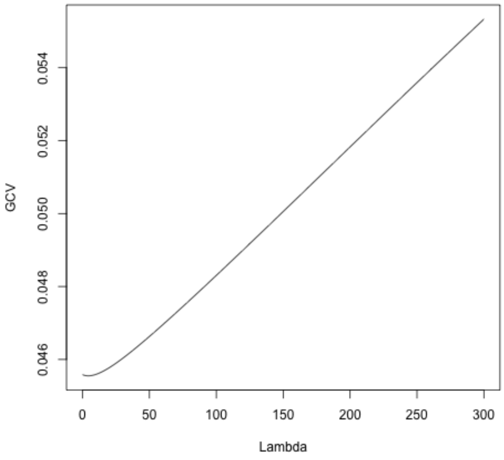

Figure 9: CV: ridge

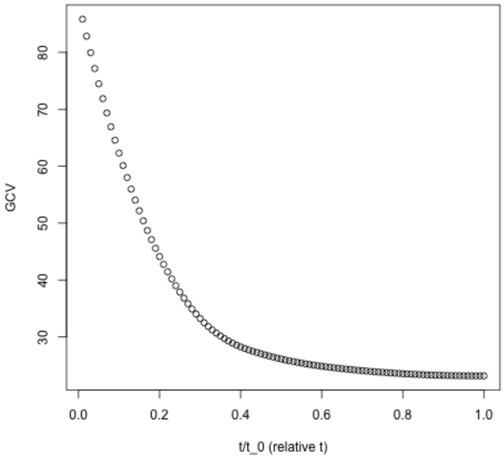

Figure 10: CV: Lasso

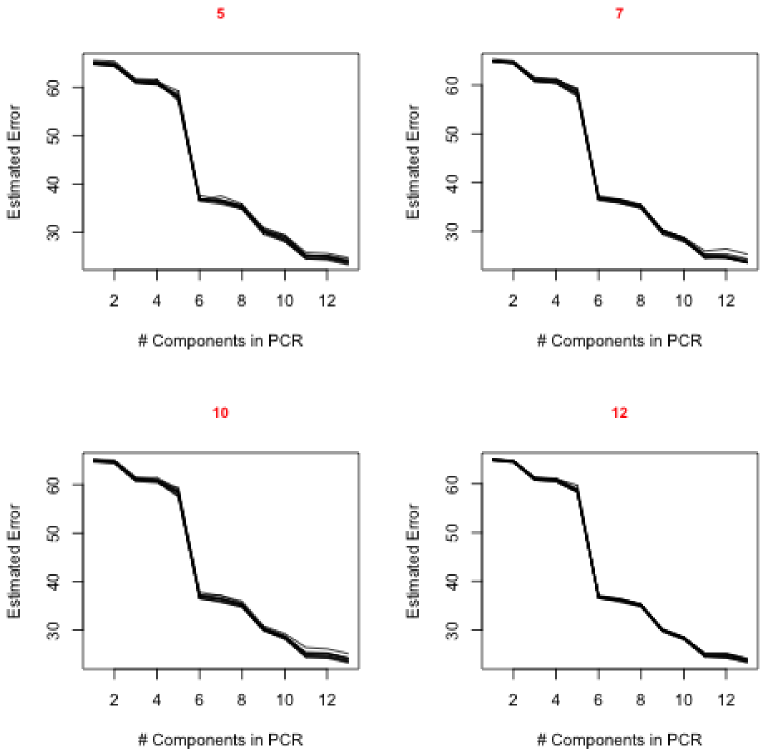

Figure 11: 10 runs CV: pcr (k=5,7,10,12)

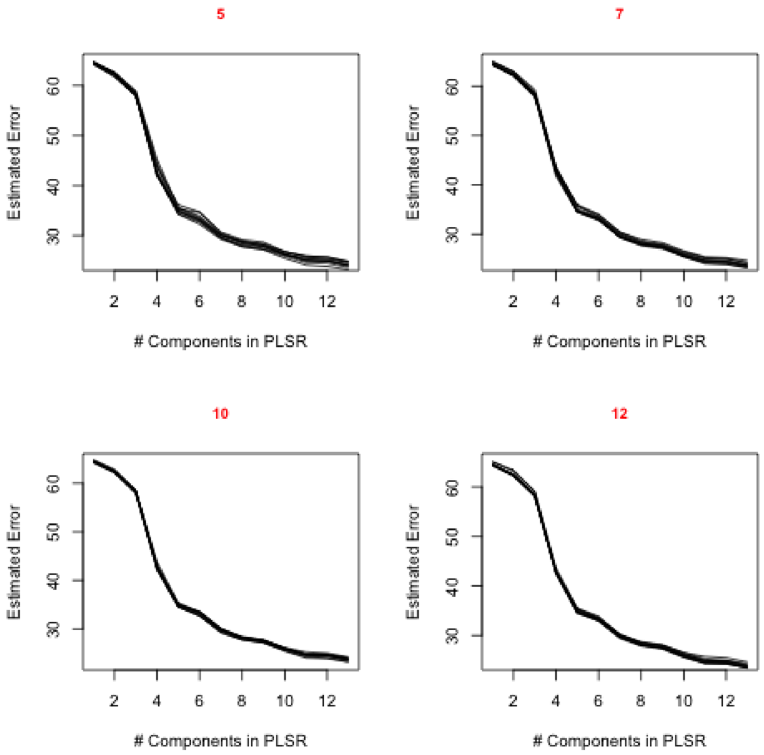

Figure 12: 10 runs CV: pls (k=5,7,10,12)