STAT 425 Case Study 2

Supanut Wanchai, Kai-Shiang Fan, Ibtisaam Dalvi, Jui-Yu Lin

1 Introduction and Data Description

This is the report for the case study of the *Crime Rates* Data Set, which is a study of crime rates, including information on 47 states of the USA in 1960. We will compare a criterion-based method using BIC as our criteria, Principal Components Analysis, and a penalized method using ridge regression to predict the crime rate. Our goal is to find the optimal model to estimate the response.

2 Data Pre-processing

Prior to doing model selection methods, we started by splitting the data set into training data(70%) and testing data(30%), along with standardizing the predictors and response (centered by their means and scaled by their standard deviations). Additionally, we factorized the categorical column *So*, which is the indicator for a southern state.

3 Model Selection Process

3.1 Criterion Based Methods: BIC

We chose BIC as our criterion-based method and got the fourth model from the Leaps and Bounds method (the *leaps* package in R, list of models generated from the method shown in Figure 1), giving us the lowest BIC. The variables corresponding to this model are *Ed* (mean years of schooling of the population aged 25 years or over), $Po1$ (per capita expenditure on police protection in 1960), *Ineq* (income inequality), and *Time*. The RMSE for the training and testing data sets are 0.4600 and 0.7438, respectively.

3.2 Principal Components Analysis

Next, we implemented the Principal Components Analysis. First of all, we looked at the Scree Plot (shown in Figure 2). A lack of an ”elbow” graph was observed, which suggests that we include all components and does not help reduce the dimensionality of predictors. So then we observed the PCR components summary table and found that the first 6 components would be sufficient to explain over 90% of the total variation of *crime rate*. The corresponding RMSE for the Training data set is 0.5380 and 0.7275 for the Testing data set. Finally, we introduced the RMSEP plot (shown in Figure 4, which suggests that selecting the first 5 components minimizes the Cross-Validation error. The corresponding RMSE for the Training data set is 0.5415 and 0.7450 for the Testing data set.

3.3 Penalized Regression: Ridge Regression

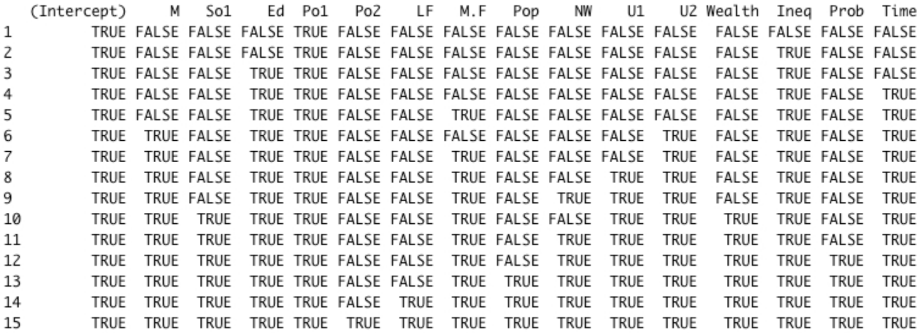
We chose ridge regression as our penalized regression and selected lambda that minimizes the Generalized Cross-Validation (GCV), shown in Figure 5. By this method, we got a model containing all the predictors with lambda = 4.5. The trace plot is shown in Figure 6, where the vertical line is the value when $\lambda = 4.5$ minimizes the GCV. The RMSE for the training and testing data sets are 0.4316 and 0.7384, respectively.

4 Conclusion

The model that gives the lowest testing error is the model from Principal Components Regression with the first 6 principal components, so we should use this model to predict *crime rate*.

5 Appendix

Methods	Training	Testing	Variables
BIC	0.4600	0.7438	Ed, Po1, Ineq, Time
PCR_5	0.5415	0.7450	The first 5 principal components
PCR_6	0.5380	0.7275	The first 6 principal components
Ridge	0.4316	0.7384	Contain all variables with $\lambda = 4.5$



```
   (Intercept)     M   So1    Ed  Po1   Po2    LF   M.F   Pop    NW    U1    U2 Wealth  Ineq  Prob  Time
1         TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE FALSE FALSE FALSE
2         TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE  TRUE FALSE FALSE
3         TRUE FALSE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE  TRUE FALSE FALSE
4         TRUE FALSE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  FALSE  TRUE FALSE  TRUE
5         TRUE FALSE FALSE  TRUE TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  FALSE  TRUE FALSE  TRUE
6         TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  FALSE  TRUE FALSE  TRUE
7         TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  FALSE  TRUE FALSE  TRUE
8         TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  FALSE  TRUE FALSE  TRUE
9         TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  FALSE  TRUE FALSE  TRUE
10        TRUE  TRUE  TRUE  TRUE TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE   TRUE  TRUE FALSE  TRUE
11        TRUE  TRUE  TRUE  TRUE TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE   TRUE  TRUE FALSE  TRUE
12        TRUE  TRUE  TRUE  TRUE TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
13        TRUE  TRUE  TRUE  TRUE TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
14        TRUE  TRUE  TRUE  TRUE TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
15        TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE  TRUE  TRUE  TRUE
```

Figure 1: Subsets of Models Generated from the Leaps and Bounds Method

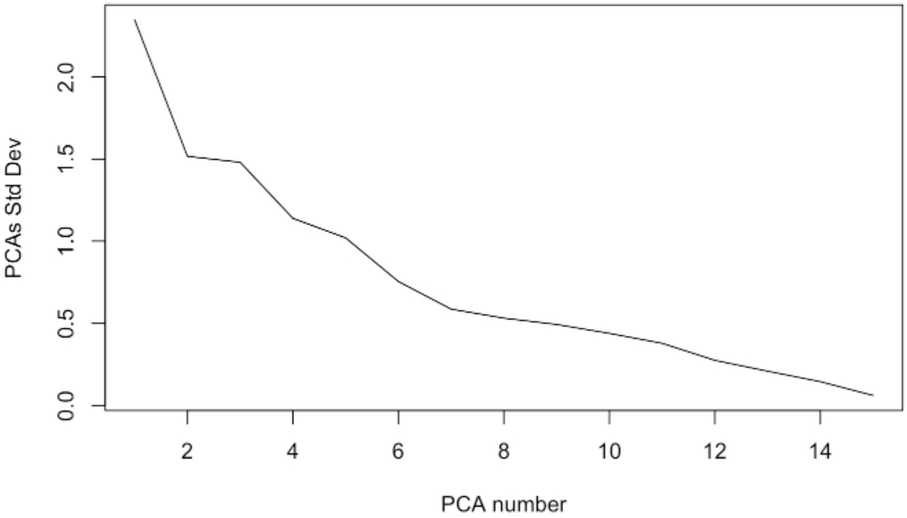
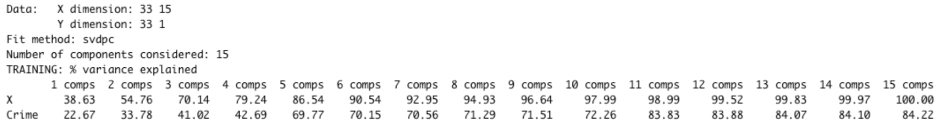


Figure 2: Scree Plot



```
Data:   X dimension: 33 15
        Y dimension: 33 1
Fit method: svdpc
Number of components considered: 15
TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X        38.63    54.76    70.14    79.24    86.54    90.54    92.95    94.93    96.64     97.99     98.99     99.52     99.83     99.97    100.00
Crime    22.67    33.78    41.02    42.69    69.77    70.15    70.56    71.29    71.51     72.26     83.83     83.88     84.07     84.10     84.22
```

Figure 3: Summary Table from PCR

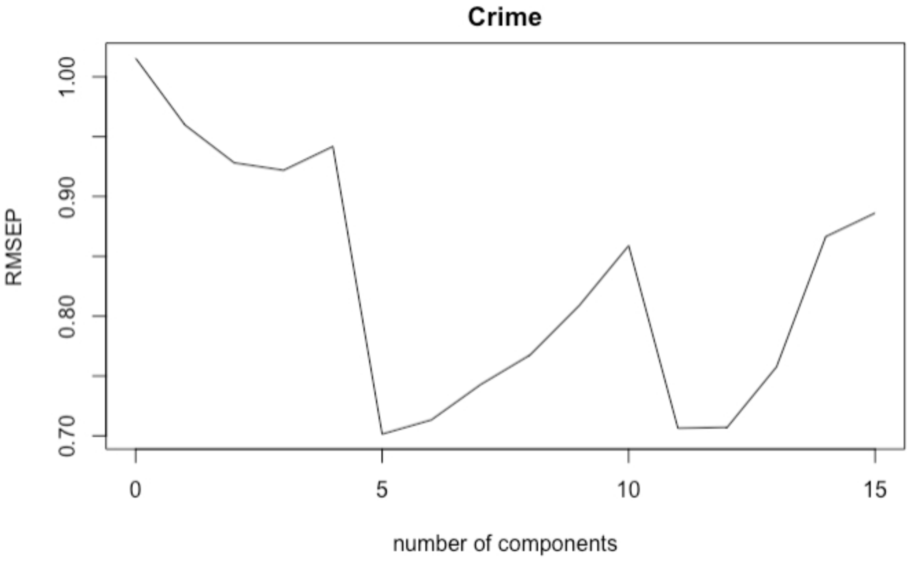
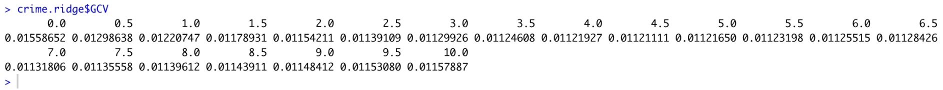


Figure 4: RMSEP Plot



```
> crime.ridge$GCV
       0.0        0.5        1.0        1.5        2.0        2.5        3.0        3.5        4.0        4.5        5.0        5.5        6.0        6.5 
0.01558652 0.01298638 0.01220747 0.01178931 0.01154211 0.01139109 0.01129926 0.01124608 0.01121927 0.01121111 0.01121650 0.01123198 0.01125515 0.01128426 
       7.0        7.5        8.0        8.5        9.0        9.5       10.0 
0.01131806 0.01135558 0.01139612 0.01143911 0.01148412 0.01153080 0.01157887 
> 
```

Figure 5: $\lambda = 4.5$ minimizes GCV

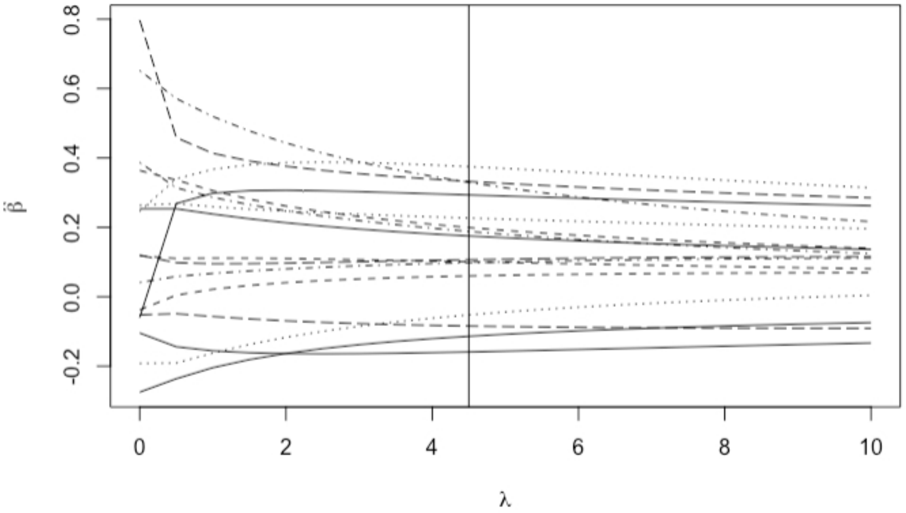


Figure 6: Trace Plot