MATH 686 Project

Daniel Krasnov

Introduction

First, the SCAD and LASSO methods are introduced for variable selection in the Cox's Proportional Hazards Model. Then results from (Fan and Li 2002) are replicated which compares these variable selection methods with AIC and BIC best subset selection. Finally a dataset is introduced to which SCAD and LASSO will be applied for my MATH 686 project.

Background

INTODUCE Survival analysis and data

INTRODUCE COX'S MODEL HERE

The following is taken from (Fan and Li 2002). Consider independent samples $(\mathbf{x}_i, Y_i)$ with conditional density $f_i(y_i; \mathbf{X}_i^T, \ )$. Let $\ell_i = \log f_i$. Then a general form of penalized likelihood is given by

$$\sum_{i=1}^{n} \ell_i y_i; \mathbf{x}_i^T \ - n \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

where d is the dimension of $\ $, $p_\lambda(.)$ is some penalty function and $|lambda$ is a tuning parameter. Selecting a function $p_\lambda(.)$ amounts to selecting a variable selection method for COx's Proportional Hazards model. In this study we will consider two penalties: LASSO and SCAD,

$$p_\lambda(|\theta|) = \lambda|\theta| \qquad \text{LASSO}$$

$$p_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \qquad \text{SCAD}$$

where $a > 2$ and $\theta > 0$. In general, a value of $a = 3.7$ is used and we adopt this for our study.

SCAD and LASSO Simulation

In this section model performance is compared on simulated data for the LASSO, SCAD, and AIC and BIC best subset regression parameter selection techniques. Model performance is assessed through the Relative Model Errors (RME)

$$\mathbb{E}\left\{\exp(-\mathbf{X}^T\hat{\beta}) - \exp(-\mathbf{X}^T\beta_0)\right\}^2.$$

We simulate 100 datasets with $n = 75$ and $n = 100$ observations from the exponential hazard model

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T\ )$$

- $\beta = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$,
- x_i are marginally standard normal with correlation $\rho = 0.5$,
- Censoring times are exponentially distributed with mean $U\exp(\mathbf{x^T}_{\mathbf{0}})$, $U \sim$ Uniform$(1, 3)$.

Our simulations yielded the following results:

```
$Initial_Beta
[1] 0.8 0.0 0.0 1.0 0.0 0.0 0.6 0.0

$Estimate_Beta_LASSO
[1] 0.69321791 0.01445390 0.04145991 0.85647956 0.02416683 0.02494662 0.49481876
[8] 0.02673219

$Estimate_Beta_SCAD
[1]  0.833263915 -0.018282435  0.014086369  1.028599102 -0.006532123
[6]  0.002125741  0.610285230  0.007945236

$Simulate_Beta_LASSO_std
[1] 0.18430473 0.10848085 0.09157499 0.19427233 0.09968270 0.11140514 0.17312590
[8] 0.11216280

$Simulate_Beta_SCAD_std
[1] 0.18005439 0.10225491 0.09562182 0.18951014 0.06530528 0.10661716 0.18683072
[8] 0.11246143

$Ave_Num_of_Zero_coeff_LASSO
[1] 2.7 0.0
```

```
$Ave_Num_of_Zero_coeff_SCAD
[1] 4.18 0.00

$MRME_LASSO
[1] 58.85458

$MRME_SCAD
[1] 52.1683
```

Method	MRME(%)	Aver. no. cor. 0 coeff.	Aver. no. incor. 0 coeff.
n=75			
SCAD	60.9554	3.68	0.01
LASSO	30.3892	2.53	0
AIC	65.8362	4.06	0.06
BIC	53.48	4.62	0.12
n=100			
SCAD	52.1683	4.18	0
LASSO	58.8546	2.7	0
AIC	69.2571	4.3	0.02
BIC	55.0899	4.73	0.03

Method	Beta1 SD	Beta4 SD	Beta7 SD
n=75			
SCAD	60.9554	3.68	0.01
LASSO	30.3892	2.53	0
AIC	65.8362	4.06	0.06
BIC	53.48	4.62	0.12
n=100			
SCAD	52.1683	4.18	0
LASSO	58.8546	2.7	0
AIC	69.2571	4.3	0.02
BIC	55.0899	4.73	0.03

Application

Next we consider the applications of LASSO, SCAD, AIC, and BIC variable selection methods on real data. We use the Mayo Clinic Primary Biliary Cholangitis (PBC) data. PBC is an

autoimmune disease which damages the liver's bile ducts leading to cirrhosis and eventually death (Therneau et al. 2000). The dataset contains 418 cases of PBC, 312 of which are from a randomized trial and 106 cases of patients not present in the trial but agreed to be tracked. The data used are available in the `survival` R package under the variable `pbc`. A table of covariates present in the data is available below (Therneau and Lumley 2015).

Table 3: Description of the Mayo Clinic Primary Biliary Cholangitis dataset.

Variable	Description
age	in years
albumin	serum albumin (g/dl)
alk.phos	alkaline phosphatase (U/liter)
ascites	presence of ascites
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirubin (mg/dl)
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
hepato	presence of hepatomegaly or enlarged liver
platelet	platelet count
protime	standardized blood clotting time
sex	m/f
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead
time	number of days between registration and the earlier of death, transplantation, or study analysis in July, 1986
trt	1/2/NA for D-penicillamine, placebo, not randomized
trig	triglycerides (mg/dl)

```
library(survival)
head(pbc)
```

```
  id time status trt      age sex ascites hepato spiders edema bili chol
1  1  400      2   1 58.76523   f       1      1       1   1.0 14.5  261
2  2 4500      0   1 56.44627   f       0      1       1   0.0  1.1  302
3  3 1012      2   1 70.07255   m       0      0       0   0.5  1.4  176
4  4 1925      2   1 54.74059   f       0      1       1   0.5  1.8  244
5  5 1504      1   2 38.10541   f       0      1       1   0.0  3.4  279
6  6 2503      2   2 66.25873   f       0      1       0   0.0  0.8  248
```

```
  albumin copper alk.phos    ast trig platelet protime stage
1    2.60    156   1718.0 137.95  172      190    12.2     4
2    4.14     54   7394.8 113.52   88      221    10.6     3
3    3.48    210    516.0  96.10   55      151    12.0     4
4    2.54     64   6121.8  60.63   92      183    10.3     4
5    3.53    143    671.0 113.15   72      136    10.9     3
6    3.98     50    944.0  93.00   63       NA    11.0     3
```

```
dim(pbc)
```

```
[1] 418  20
```

References

Fan, Jianqing, and Runze Li. 2002. "Variable Selection for Cox's Proportional Hazards Model and Frailty Model." *The Annals of Statistics* 30 (1): 74–99.

Therneau, Terry M, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. 2000. *The Cox Model*. Springer.

Therneau, Terry M, and Thomas Lumley. 2015. "Package 'Survival'." *R Top Doc* 128 (10): 28–33.