Generalized Ridge & Lasso Regression

Readings ISLR 6, Casella & Park

STA 521 Duke University

Merlise Clyde

March 20, 2017

Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ $\mathbf{X}$ is centered and scaled predictors
- (Classical) Ridge Regression controls how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\beta)$$

  subject to

$$\sum \beta_j^2 \leq t$$

- Penalized Likelihood

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2$$

- Bayesian Ridge Regression - Hierarchical prior
  - $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$
  - $\boldsymbol{\beta} \mid \phi, \kappa \sim \mathrm{N}(\mathbf{0}, \mathbf{I}(\phi\kappa)^{-1})$
  - prior on κ

For fixed κ the Bayes MAP and the penalized MLE are the same

Differences

Treatement of uncertainty

- Frequentist: use of cross validation or optimization for finding k
- Bayes: removes "nuisance" parameter κ through integration rather than than optimization
  - Can use full posterior distribution for credible intervals for parameters, regression function or predictions
  - Other Choices of priors?

Lasso

Tibshirani (JRSS B 1996) proposed estimating coefficients through L_1 constrained least squares "Least Absolute Shrinkage and Selection Operator"

- Control how large coefficients may grow

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\boldsymbol{\beta})$$

  subject to

$$\sum |\beta_j| \leq t$$

- Equivalent Quadratic Programming Problem for "penalized" Likelihood

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{1}\bar{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- Posterior mode

$$\max_{\boldsymbol{\beta}} -\{\|\mathbf{Y} - \mathbf{1}\beta_0 - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1\}$$

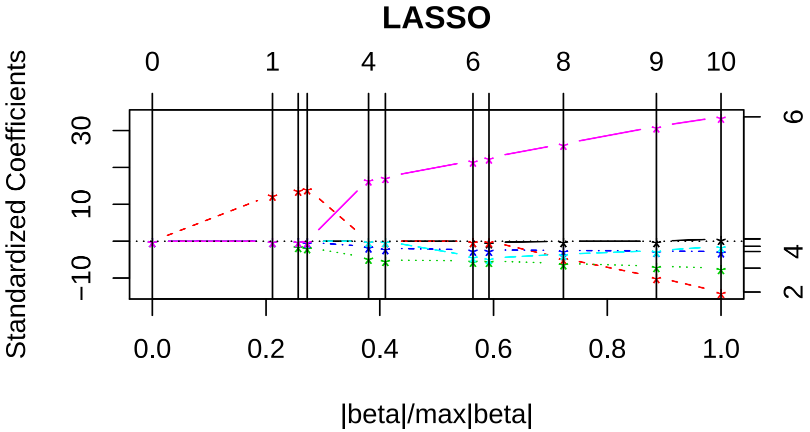
Variable Selection via the LASSO

$p = 2$ constraint $|\beta_1| + |\beta_2| \leq t$ is a diamond

R Code

Path of solutions can be found using the "Least Angle Regression" Algorithm of Efron et al (Annals of Statistics 2004)

```
library(lars) longley.lars = lars(as.matrix(longley[,-7]),
longley[,7], type="lasso") plot(longley.lars)
```



Solutions

```r
kable(coef(longley.lars), digits=4)
```

GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0327	0.0000	0.0000	0.0000	0.0000
0.0000	0.0362	-0.0037	0.0000	0.0000	0.0000
0.0000	0.0372	-0.0046	-0.0010	0.0000	0.0000
0.0000	0.0000	-0.0124	-0.0054	0.0000	0.9068
0.0000	0.0000	-0.0141	-0.0071	0.0000	0.9438
0.0000	0.0000	-0.0147	-0.0086	-0.1534	1.1843
-0.0077	0.0000	-0.0148	-0.0087	-0.1708	1.2289
0.0000	-0.0121	-0.0166	-0.0093	-0.1303	1.4319
0.0000	-0.0253	-0.0187	-0.0099	-0.0951	1.6865
0.0151	-0.0358	-0.0202	-0.0103	-0.0511	1.8292

Which one?

Summary

```
sum.lars = summary(longley.lars)
sum.lars

## LARS/LASSO
## Call: lars(x = as.matrix(longley[, -7]), y = longley[, 7], ty
##    Df     Rss        Cp
## 0   1 185.009 1976.7120
## 1   2   6.642   59.4712
## 2   3   3.883   31.7832
## 3   4   3.468   29.3165
## 4   5   1.563   10.8183
## 5   4   1.339    6.4068
## 6   5   1.024    5.0186
## 7   6   0.998    6.7388
## 8   7   0.907    7.7615
## 9   6   0.847    5.1128
## 10  7   0.836    7.0000
```

Cp Solution

Min $C_p = SSE_p/\hat{\sigma}^2_F - n + 2p$
For a model that includes all true predictors $C_p \approx p$

```
n.sol = length(sum.lars$Cp)
best = which.min(abs(sum.lars$Cp - sum.lars$Df)[-n.sol])
kable(coef(longley.lars)[best,], digits=4)
```

GNP.deflator	0.0000
GNP	0.0000
Unemployed	-0.0147
Armed.Forces	-0.0086
Population	-0.1534
Year	1.1843

Can also use Cross-Validation - many packages available!
What about uncertainty? Confidence intervals?

Bayesian Lasso

Park & Casella (JASA 2008) and Hans (Biometrika 2010) propose Bayesian versions of the Lasso

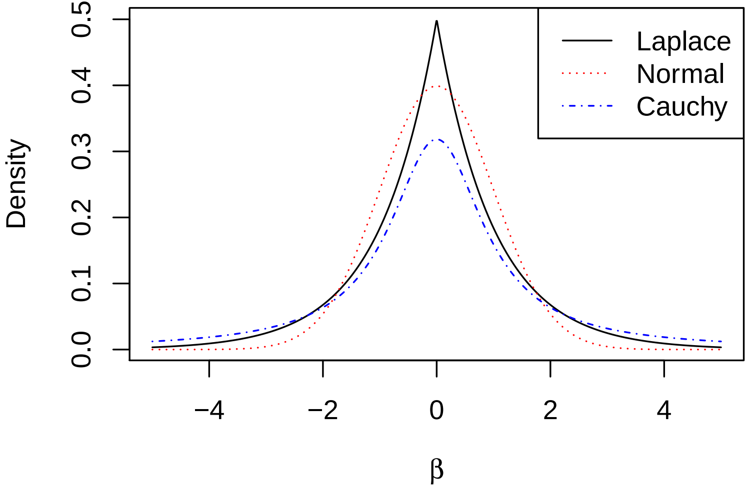
$$\begin{aligned}
\mathbf{Y} \mid \beta_0, \boldsymbol{\beta}, \phi &\sim \mathsf{N}(\mathbf{1}_n \beta_0 + \mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n/\phi) \\
\boldsymbol{\beta} \mid \beta_0, \phi, \boldsymbol{\tau} &\sim \mathsf{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\tau}^2)/\phi) \\
\tau_1^2 \ldots, \tau_p^2 \mid \beta_0, \phi &\overset{\text{iid}}{\sim} \mathsf{Exp}(\lambda^2/2) \\
p(\beta_0, \phi) &\propto 1/\phi
\end{aligned}$$

Can show that $\beta_j \mid \phi, \lambda \overset{\text{iid}}{\sim} DE(\lambda\sqrt{\phi})$

$$\int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2}\phi\frac{\beta^2}{s}} \, \frac{\lambda^2}{2} e^{-\frac{\lambda^2 s}{2}} \, ds = \frac{\lambda\phi^{1/2}}{2} e^{-\lambda\phi^{1/2}|\beta|}$$

Scale Mixture of Normals (Andrews and Mallows 1974)

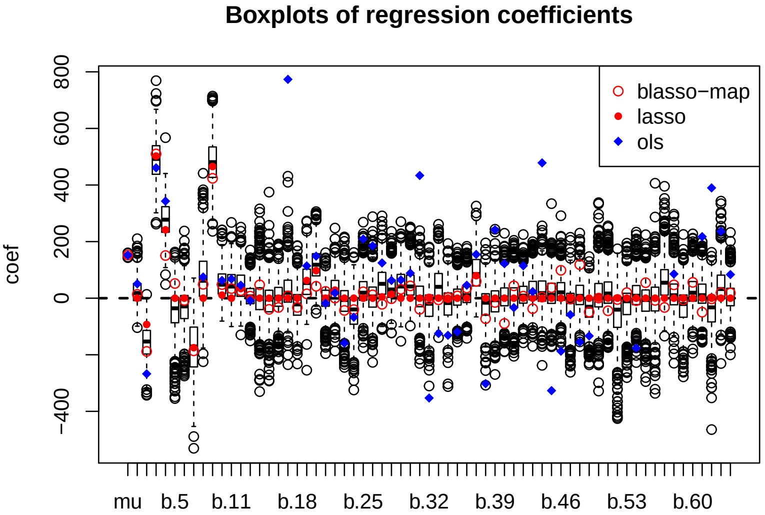
Densities



Bayesian Lasso Fitting

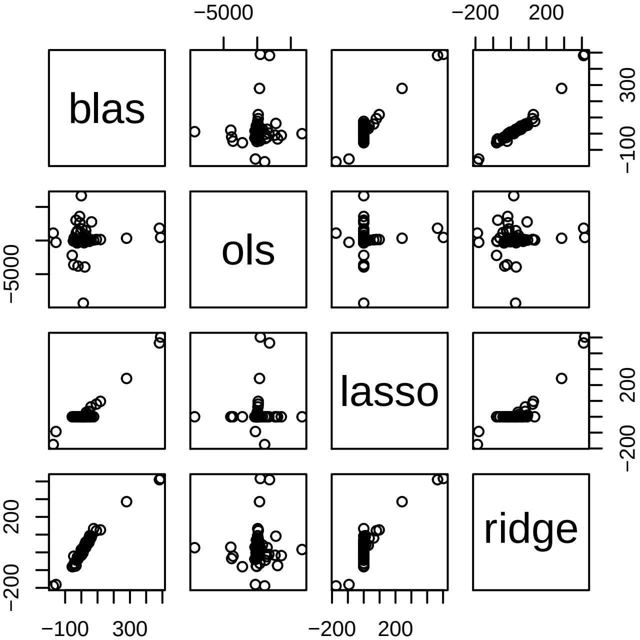
```r
data(diabetes)
Y = diabetes$y
X = diabetes$x2  # 64 variables from all 10 main effects,
                  # two-way interactions and quadradics
set.seed(8675309)
suppressMessages(library(monomvn))

## Ordinary Least Squares regression from monomvn
reg.ols <- regress(X, Y)
## ridge regression
reg.ridge <- regress(X, Y, method="ridge")
## Lasso regression from monomvn
reg.las <- regress(X, Y, method="lasso")

## Bayesian Lasso regression from monomvn
reg.blas <- blasso(X, Y, RJ=FALSE, verb=0)
```

Estimates



Boxplots of regression coefficients
coef
800
600
400
200
0
−400
mu
b.5
b.11
b.18
b.25
b.32
b.39
b.46
b.53
b.60
blasso−map
lasso
ols

Shrinkage



blas
ols
lasso
ridge
−5000
−200
200
300
−100
−5000
200
−200
200
−200
−100
300
−200
200

Summary

- Bayesian and Regular LASSO shrink (unstable) coefficients to zero
- Bayesian posterior mean cannont be zero (so no selection)
- Bayesian MAP (Maximum a posteriori) estimate equivalent to Lasso penalized MLE for same λ
- Bayesian allows uncertainty in λ to propogate to estimates and predictions
- Bayesian MAP estimates via EM algorithms or Variational Bayes (STAN)
- Report MAP estimate and HPD intervals
- RJ = TRUE incorporates probabilty that $\beta = 0$ for variable selection