Lasso and Bayesian Lasso Regression

Readings ISLR 6, Casella & Park

STA 521 Duke University

Merlise Clyde

October 30, 2019

Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathbf{X}$ is centered and scaled predictors

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

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

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

subject to

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

quadratic with constraints

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

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

subject to

$$\sum \beta_j^2 \leq t \quad \text{(P1)}$$

related via lagrange multiplier

- Penalized Likelihood

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2 \quad \text{(P2)}, \quad k \geq 0$$

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

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

subject to

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

Freq. proced.

- 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

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

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

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
  - $p(\beta_0, \phi \mid \boldsymbol{\beta}, \kappa) \propto \phi^{-1}$ ← prior on precision

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

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

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 shrink on to β zero

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

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

subject to

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

opt

- Penalized Likelihood

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

opt

- 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

prior on κ

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

Differences

Treatement of uncertainty

Differences

Treatement of uncertainty

- Frequentist: use of cross validation or optimization for finding k

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, Bayes - marginalize out "nuisance" parameters

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

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"

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

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

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

subject to

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

Quadratic in loss

linear in weights

why does the l_1 penalty make a difference?

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

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

subject to

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

(P1)

Lagrange multipliers

- Equivalent Quadratic Programming Problem for "penalized" Likelihood

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

(P2)

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

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

subject to

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

- Equivalent Quadratic Programming Problem for "penalized" Likelihood

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

this is not an algorithm for finding $\hat{\beta}$

minimization prob.

- Posterior mode

$$\max_{\beta} \left[Post(\beta | Y, X) \right]$$

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

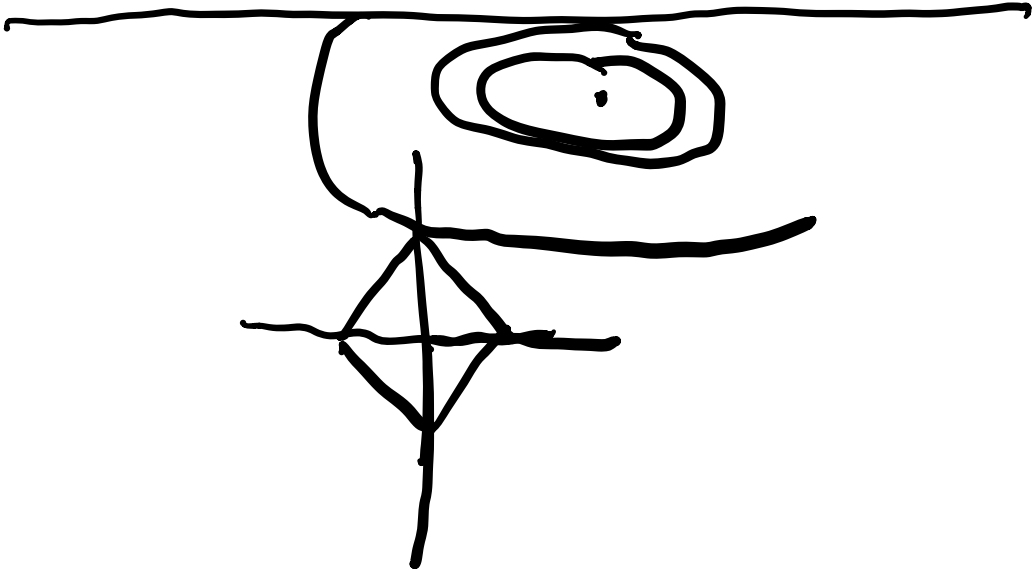
Variable Selection via the LASSO

$$\mathcal{L}ik(Y, X; \beta) = \exp\left(-\|Y - \bar{Y}\mathbf{1} - X\beta\|^2\right)$$

$$\pi(\beta) \propto \exp\left(-\lambda \|\beta\|_1\right)$$

$$\|\beta\|_1 = \sum |\beta_i|$$

$p = 2$ 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)

(1) Provide an algorithm for finding $\hat{\beta}_\lambda$ for $\lambda \geq 0$

$$\frac{1}{n}\sum_i (y_i - \bar{y} - \beta^T x_i)^2 + \lambda \|\beta\|_1$$

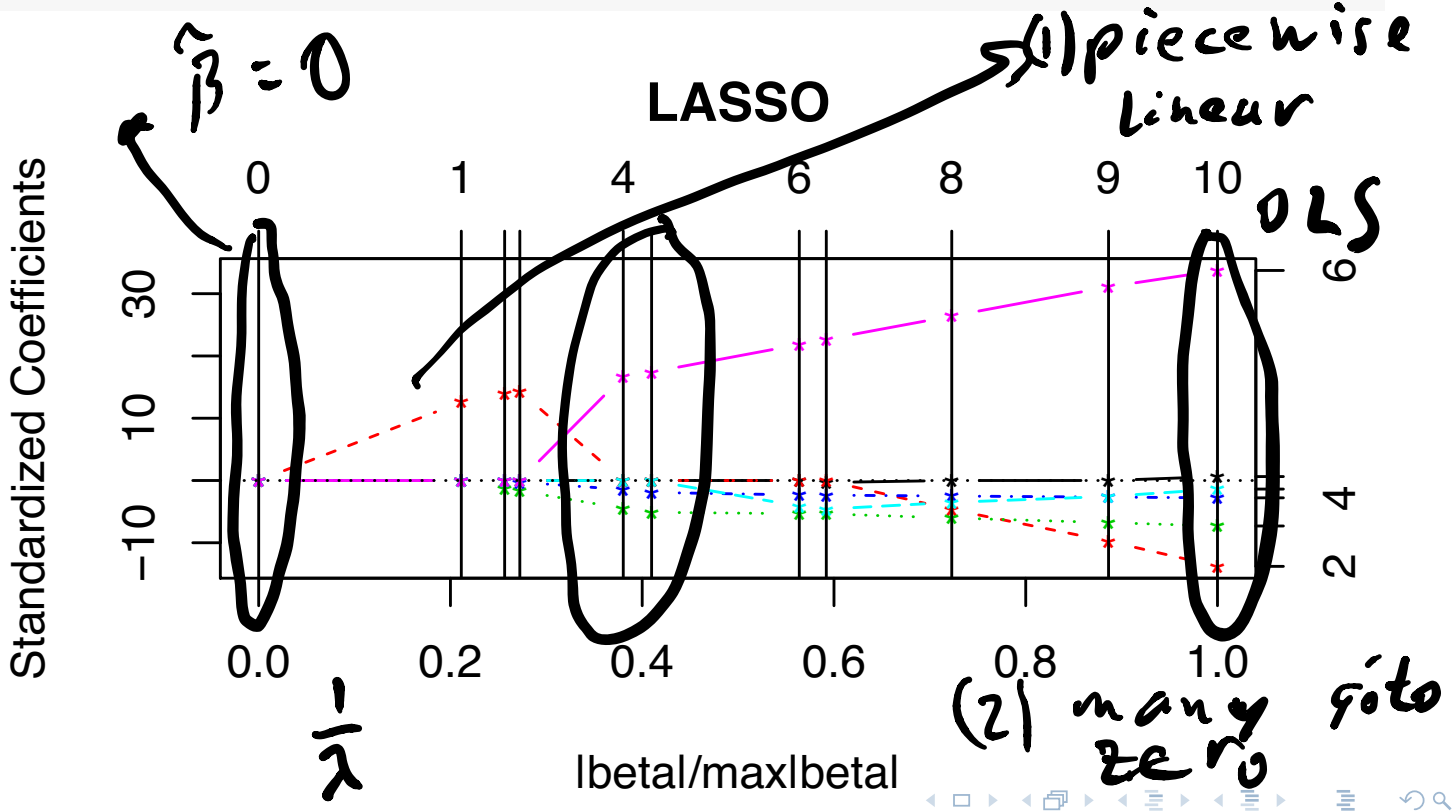
(2) Provide an algorithm for finding $\hat{\beta}_\lambda$ for all $\lambda \geq 0$

regularization path or path of solutions

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

```
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 $C_p = SSE_p/\hat{\sigma}_F^2 - n + 2p$

Cp Solution

one model selection criteria

Min $C_p = SSE_p/\hat{\sigma}_F^2 - n + 2p$
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)
```

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

really just do cross-validation

Cp Solution

Min $C_p = SSE_p/\hat{\sigma}_F^2 - n + 2p$
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)
```

	x
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!

Cp Solution

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

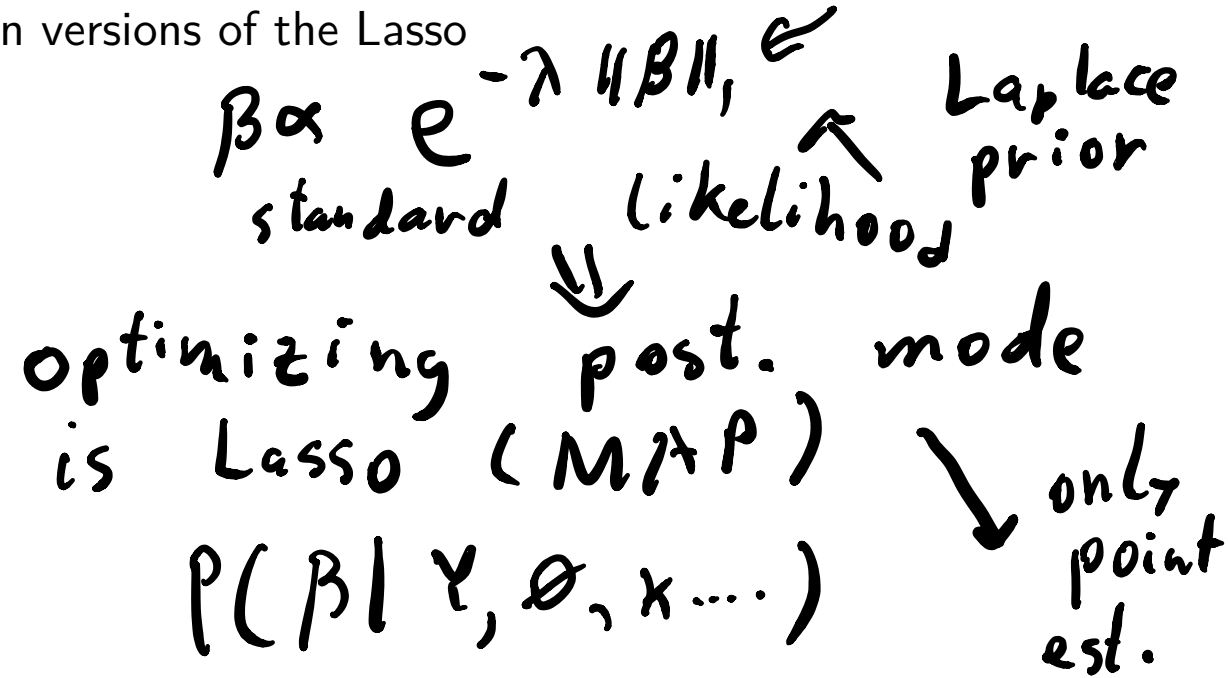
```
n.sol = length(sum.lars$Cp)
best = which.min(abs(sum.lars$Cp - sum.lars$Df)[-n.sol])
kable(coef(longley.lars)[best,], digits=4)
```

	x
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



Bayesian Lasso

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

$$\mathbf{Y} \mid \beta_0, \boldsymbol{\beta}, \phi \quad \sim \quad \mathrm{N}(\mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n/\phi)$$

ϕ prec. of noise or error

Bayesian Lasso

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

$$\begin{aligned} \mathbf{Y} \mid \beta_0, \boldsymbol{\beta}, \phi &\sim \mathrm{N}(\mathbf{1}_n\beta_0 + \mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n/\phi) \\ \boldsymbol{\beta} \mid \beta_0, \phi, \boldsymbol{\tau} &\sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\tau}^2)/\phi) \end{aligned}$$

β looks normal cond. of $\tau = \tau_1 \cdots \tau_p$

Bayesian Lasso

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

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

ϕ is global not coordinate wise based

τ_i^2 is local: is coordinate specific

so we have a notion of scale on each coord. which is variable

Bayesian Lasso

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

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

Bayesian Lasso

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

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

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

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

Bayesian Lasso

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

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

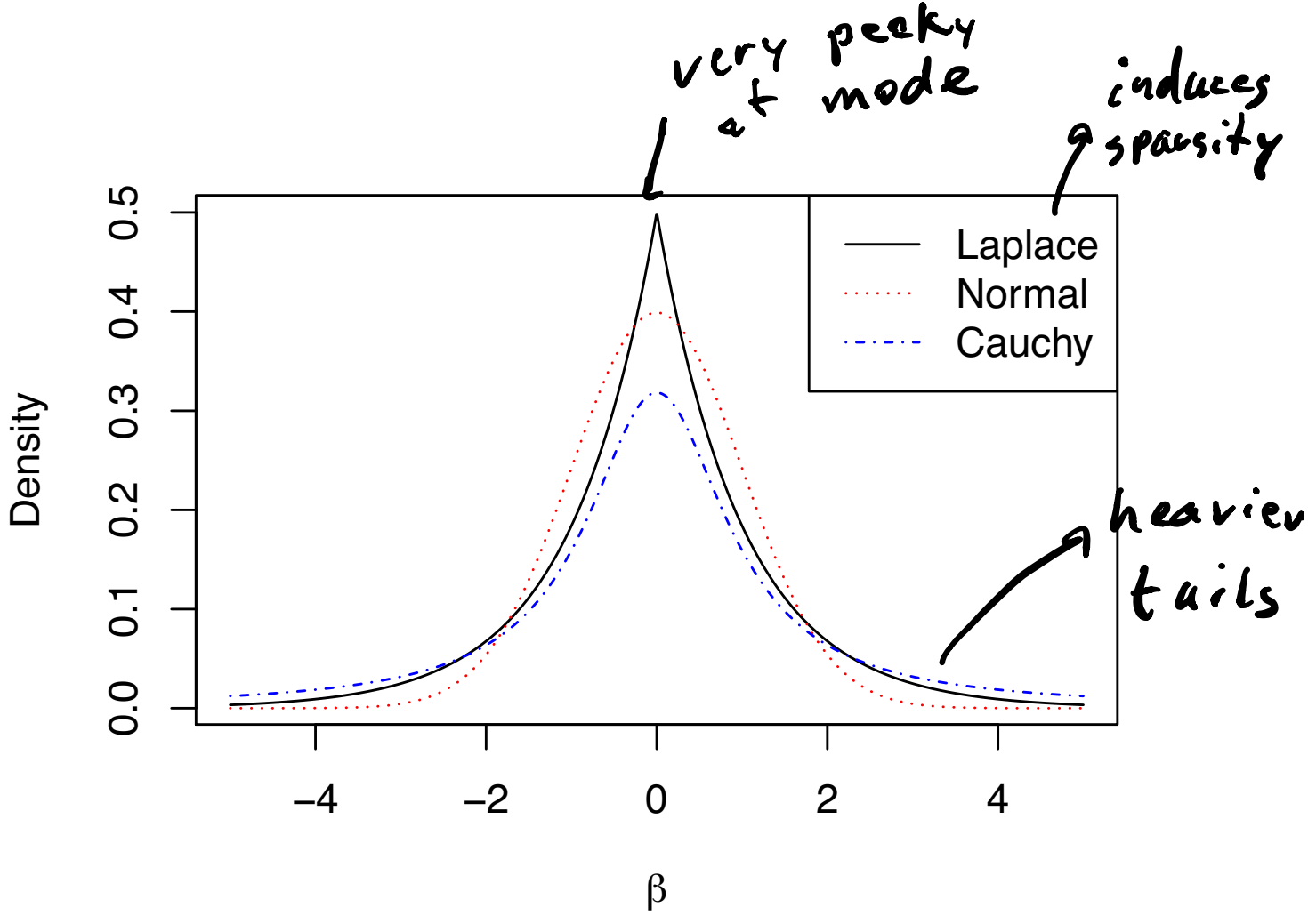
Bayesian Lasso

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)

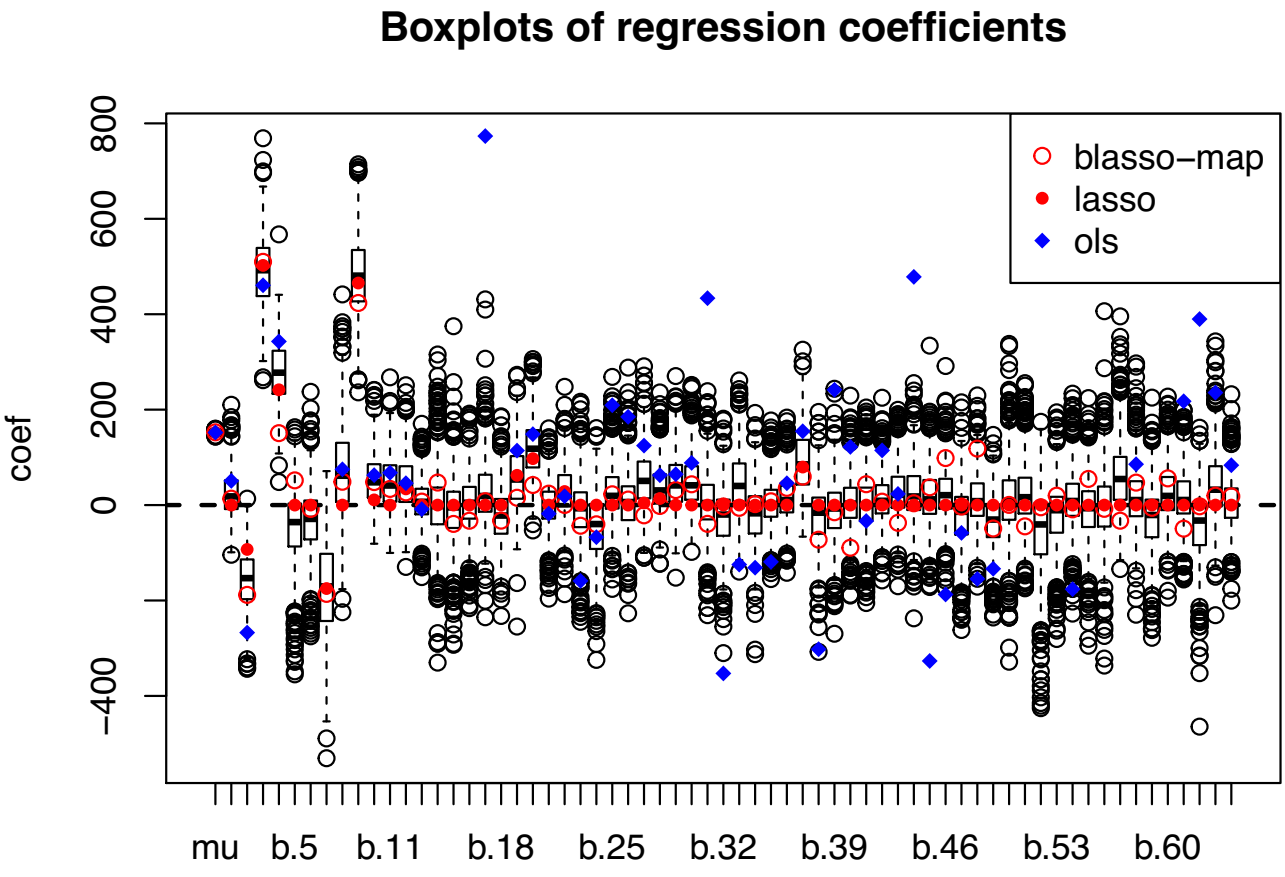
Densities


very peaky at mode
induces sparsity
heavier tails
Laplace
Normal
Cauchy
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
β

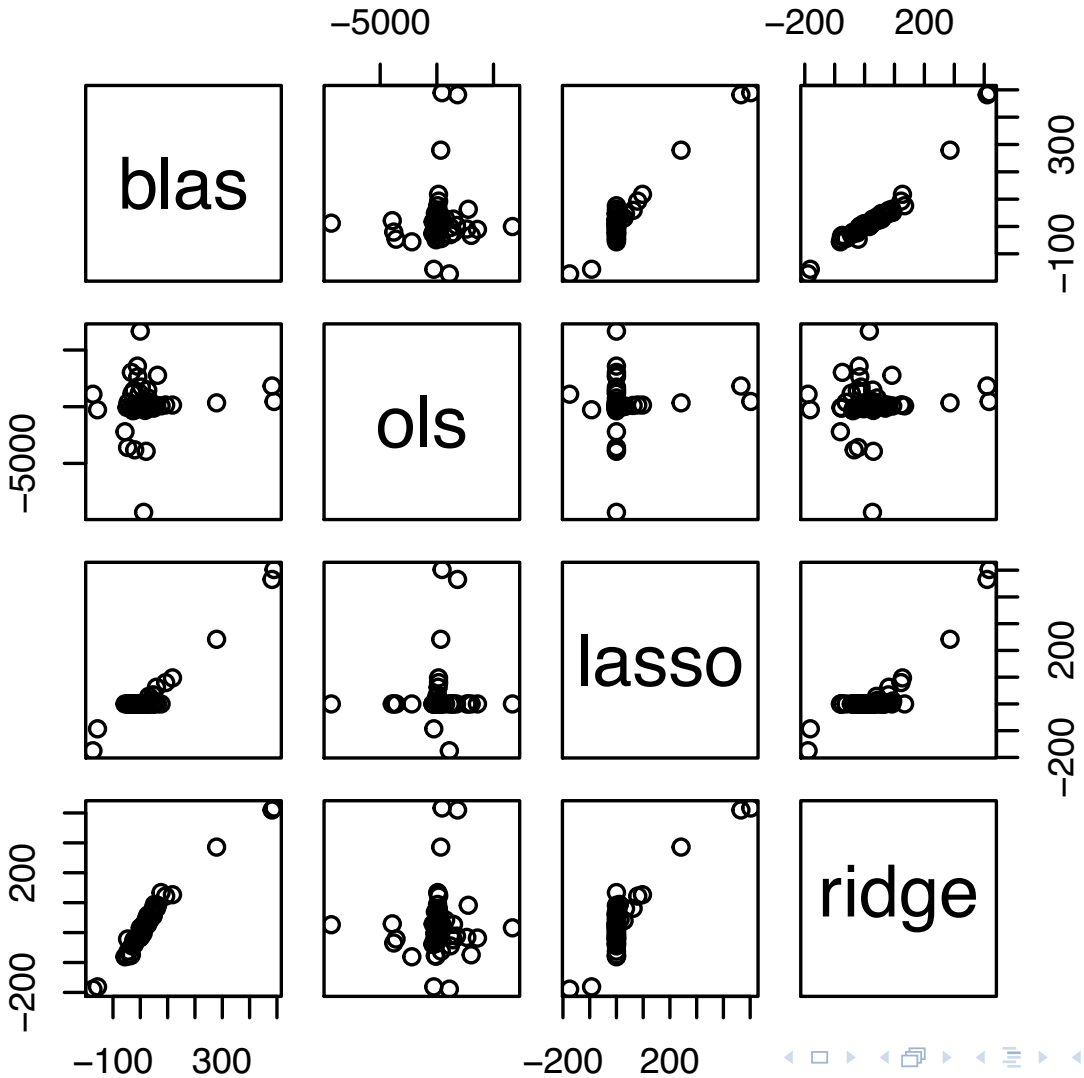
Bayesian Lasso Fitting

```
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
blasso−map
lasso
ols
mu
b.5
b.11
b.18
b.25
b.32
b.39
b.46
b.53
b.60

Shrinkage


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

Summary

- Bayesian and Regular LASSO shrink (unstable) coefficients to zero

Summary

- Bayesian and Regular LASSO shrink (unstable) coefficients to zero
- Bayesian posterior mean cannont be zero (so no selection)

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

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

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)

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

Algo.

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

LECTURE 10
Sparse regression

We have seen previously that for the case that $p \gg n$ the following ridge regression model allows us stable inference

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|^2, \quad \lambda > 0.$$

Often we don't just want a good predictive model but we also want to know which variables are relevant to the prediction. The problem of simultaneously inferring a good regression model as well as selecting variable is called simultaneous regression and variable selection. In this lecture we will state some standard methods for simultaneous regression and variable selection.

We fist state the standard model

$$Y_i = (\beta^*)^T x_i + \varepsilon_i,$$

however we now assume that the regression coefficients are zero for the majority coordinates ($i = 1, .., p$). The subset of non-zero coordinates for the true model $\mathcal{A}_* = \{j : |\beta_*^{(j)}| \neq 0\}$ and the number of non-zero coefficients is denoted as $|\mathcal{A}_*|$. Our objective is given data $D = \{(x_i, y_i)_{i=1}^n\}$ to infer $\widehat{\beta}$ such that

(1) Selection consistency: The non-zero subset of $\widehat{\beta}$ is denoted as $\widehat{\mathcal{A}} = \{j : |\widehat{\beta}^{(j)}| \neq 0\}$. We would like the two subsets $\mathcal{A}_*$ and $\widehat{\mathcal{A}}$ to be close for any finite n and identical as $n \to \infty$.

(2) Estimation consistency: How well do the coefficients in the selected set converge:

$$\lim_{n \to \infty} \widehat{\beta}_{\mathcal{A}_*} = \beta^*_{\mathcal{A}_*}$$

The approach we will use for simultaneous regression and variable selection is the following minimization problem

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_q^q, \quad \lambda > 0,$$

where $\|\beta\|_q^q$ is a penalization by the q-norm. We've already seen the result of minimizing the 2-norm leads to ridge regression. We will now explore two other norms: the 1-norm and the 0-norm.

We start with the zero norm

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_0^0, \quad \lambda > 0,$$

This is equivalent to the following minimization problem

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} I(\beta_j \neq 0), \quad \lambda > 0,$$

which is suggesting minimizing the square error using the fewest variables possible with λ acting as the tradeoff between the number of variables and the error. The above minimization problem is NP-hard as it reduces to to exact cover by three sets. This means we can't practically implement the above optimization problem with any efficiency, even $p = 10$ requires a search over a massive space.

10.1. LASSO: Least Absolute Selection and Shrinkage Operator

The idea behind the lasso procedure is to minimize

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1, \quad \lambda > 0,$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. For reasons we will discuss minimizing the above penalized loss function results in variable selection and regression, results in regression coefficients that are exactly zero. An argument has been made that minimizing the 1-norm regularized problem is a good approximation of the 0-norm minimization problem. We will explore both why this minimization problem approximates the 0-norm as well procedures to minimize the 1-norm.

10.1.1. The geometry of polytopes

Recall that there is an equivalence between

$$\arg\min_{\beta \in \mathbb{R}^p} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1,$$
$$\arg\min_{\beta \in \mathbb{R}^p} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \text{ subject to } \|\beta\|_1 \leq \tau.$$

We will contrast the following two minimization problems

$$\arg\min_{\beta \in \mathbb{R}^p} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \text{ subject to } \|\beta\|_1 \leq \tau.$$
$$\arg\min_{\beta \in \mathbb{R}^p} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \text{ subject to } \|\beta\|_2^2 \leq \tau.$$

Solutions to the upper problem is constrained to a 1-norm ball around the origin and the solutions to the lower problem is constrained to a 2-norm around the origin. Consider the true β vector to be β_*, the geometry of the square loss has ellipses as contours of equal loss. The minimizer is the smallest loss value that intersects the boundary of the p-norm ball.In the figure below we show this for two variables.

A minimizer with a sparse solution will touch/intersect the contours of the error ellipses on the axes that is sparse faces of the p-dimensional polytope. For example, when the constraint is the 2-norm ball around the origin it is very unlikely that the intersecting point will be concentrated on the axes. The geometry of the 1-norm ball especially in high dimensions intersects the ellipse at a few points. For example, the 0-norm is a star or spike that is on the axes so it will always be sparse.

Although we have considered the constrained optimization 1-norm problem the same results hold for lasso.

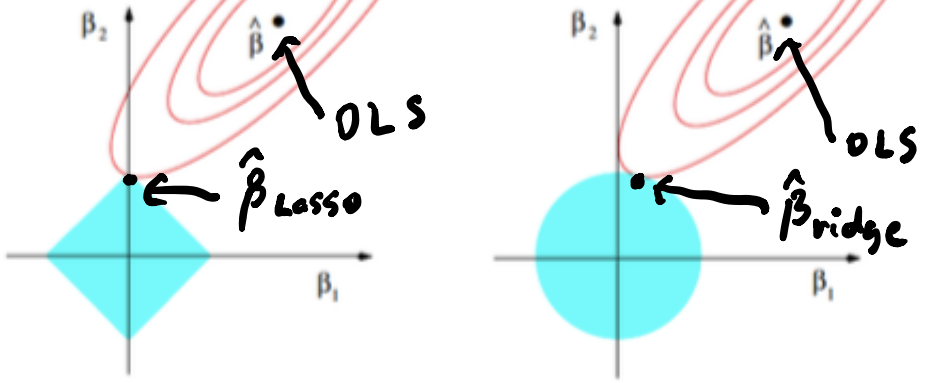


Figure 1. The 1-norm minimization for two variables is on the left and the 2-norm minimization is on the right.

10.1.2. The regularization path

Recall the optimization problem

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1, \quad \lambda > 0,$$

It is known that for $\lambda = 0$ the solution $\widehat{\beta}_{\text{LASSO}} = \widehat{\beta}_{\text{OLS}}$ and for $\lambda = \infty$ the solution is $\widehat{\beta}_{\text{LASSO}} = 0$. The regression coefficients β_λ for the lasso with regularization parameter λ is a p-dimensional vector with many of its values set to zero for larger values of λ. The idea of the regularization path is to examine how the β's change with λ the picture one should consider is λ as the x-axis and the β's on the y-axis. It is a mathematical fact that the graph of the β's will be piecewise continuous and approach zero at some point.

The idea behind the regularization path is to help select how many variables to keep in the model. In the ridge model it is hard to interpret a regularization parameter as coefficients are not sent to zero and the changes are slow. This is somewhat mitigated in the lasso model.

In the figure below we consider two regression analyses, one using ridge and the other with lasso, the dataset is a prostate cancer related problem. The response variable is PSA, a biomarker marker for prostate cancer, the covariates are several other biomarkers.

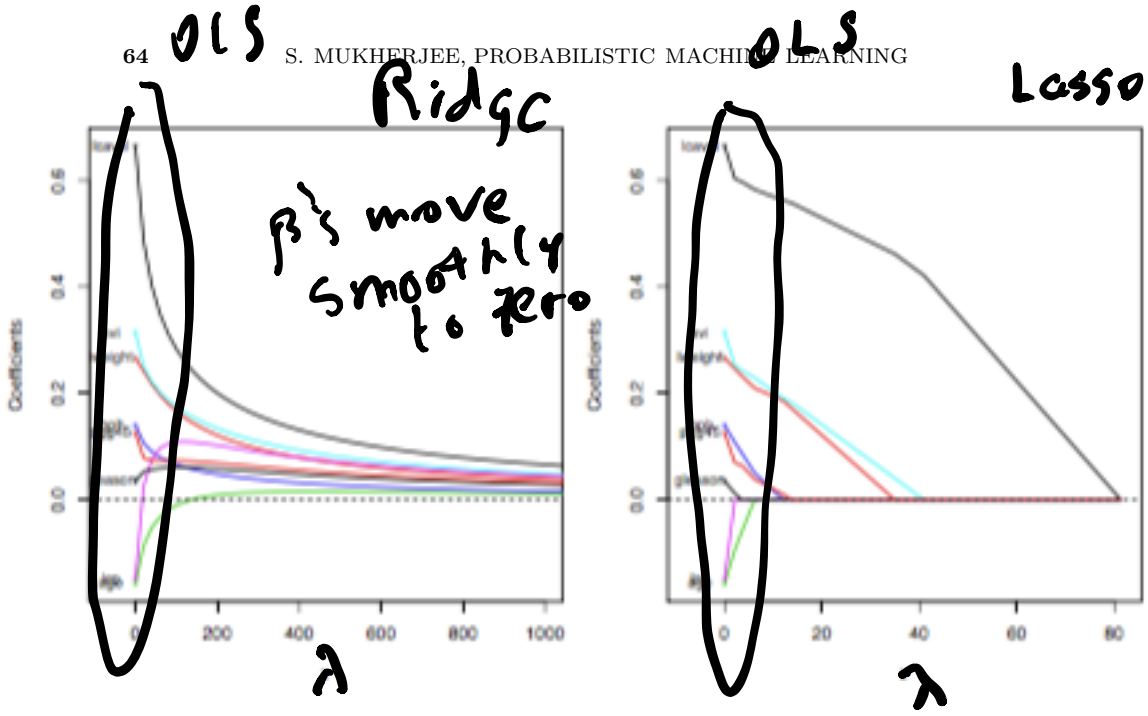


Figure 2. The left figure is the regularization path for ridge regression, the x-axis is the λ parameter and the y-axis is plotting the coefficients. The right figure is the same plot but for lasso. The response variable is PSA, a biomarker marker for prostate cancer, the covariates are a several biomarkers.

Elastic net: $\lambda > 0, \alpha \in [0,1]$

$$\min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2} + \lambda\left[\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1\right]$$

$\alpha = 1 \Rightarrow$ Ridge

$\alpha = 0 \Rightarrow$ Lasso

λ, α

$\alpha = .5$

Bayesian version of EL

1) Lid over α

1) dist. over α

2) heirarchical model

3) marginalize out mixing dist. over α