Robust Regression and Related Methods

Readings ISLR Chapter 6 + Papers

STA521 Predictive Models Duke University

Merlise Clyde

March 27, 2017

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior
$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$
  - hard thresholding to zero

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$
  - hard thresholding to zero
  - soft thresholding of non-zero coefficients to zero

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$
  - hard thresholding to zero
  - soft thresholding of non-zero coefficients to zero
- cannot achieve an optimal balance of both

Problem with Lasso/Bayesian Lasso Model

- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$
  - hard thresholding to zero
  - soft thresholding of non-zero coefficients to zero
- cannot achieve an optimal balance of both
- Carvhalo, Polson, Scott proposed the Horseshoe prior to address this problem

Problem with Lasso/Bayesian Lasso Model

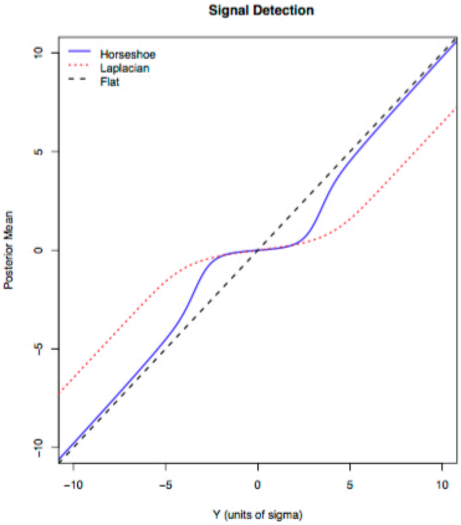
- Model: $\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$; $\mathbf{X}$ is matrix of centered and scaled predictors so that $\text{diag}(\mathbf{X}^T\mathbf{X}) = \mathbf{I}_p$
- Lasso Prior

$$\beta_j \mid \phi \overset{\text{iid}}{\sim} DE(\phi^{1/2}\lambda)$$

$$p(\beta_0, \phi) \propto 1/\phi$$

- One parameter λ controls shrinkage of $\boldsymbol{\beta}$
  - hard thresholding to zero
  - soft thresholding of non-zero coefficients to zero
- cannot achieve an optimal balance of both
- Carvhalo, Polson, Scott proposed the Horseshoe prior to address this problem

Robust Shrinkage



http://www.jmlr.org/proceedings/papers/v5/carvalho09a/carvalho09a.pdf

Horseshoe

- Horseshoe Prior Distribution

$$\beta_j \mid \phi, \tau_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \tau_j/\phi)$$

Horseshoe

- Horseshoe Prior Distribution

$$\beta_j \mid \phi, \tau_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \tau_j/\phi)$$

- $\tau_j^{1/2} \mid \varphi \overset{\text{iid}}{\sim} \mathsf{C}^+(0, \varphi^{1/2})$ with density for the half-Cauchy

$$p(\tau^{1/2}) \propto \left(1 + \frac{\tau}{\varphi}\right)^{-1} \qquad \tau > 0$$

Horseshoe

- Horseshoe Prior Distribution

$$\beta_j \mid \phi, \tau_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \tau_j/\phi)$$

- $\tau_j^{1/2} \mid \varphi \overset{\text{iid}}{\sim} \mathsf{C}^+(0, \varphi^{1/2})$ with density for the half-Cauchy

$$p(\tau^{1/2}) \propto \left(1 + \frac{\tau}{\varphi}\right)^{-1} \qquad \tau > 0$$

- $\varphi^{1/2} \sim \mathsf{C}^+(0, 1)$

Horseshoe

- Horseshoe Prior Distribution

$$\beta_j \mid \phi, \tau_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \tau_j/\phi)$$

- $\tau_j^{1/2} \mid \varphi \overset{\text{iid}}{\sim} \mathsf{C}^+(0, \varphi^{1/2})$ with density for the half-Cauchy

$$p(\tau^{1/2}) \propto \left(1 + \frac{\tau}{\varphi}\right)^{-1} \qquad \tau > 0$$

- $\varphi^{1/2} \sim \mathsf{C}^+(0, 1)$
- $p(\beta_0, \phi) \propto 1/\phi$

Horseshoe

- Horseshoe Prior Distribution

$$\beta_j \mid \phi, \tau_j \overset{\text{iid}}{\sim} \mathsf{N}(0, \tau_j/\phi)$$

- $\tau_j^{1/2} \mid \varphi \overset{\text{iid}}{\sim} \mathsf{C}^+(0, \varphi^{1/2})$ with density for the half-Cauchy

$$p(\tau^{1/2}) \propto \left(1 + \frac{\tau}{\varphi}\right)^{-1} \qquad \tau > 0$$

- $\varphi^{1/2} \sim \mathsf{C}^+(0, 1)$
- $p(\beta_0, \phi) \propto 1/\phi$

Alternative representation

Normal + Generalized Beta:

$$\beta_j \mid \rho_j \sim N(0, 1/\rho_j - 1)$$

Alternative representation

Normal + Generalized Beta:

$$\beta_j \mid \rho_j \sim N(0, 1/\rho_j - 1)$$

$$p(\rho_j) \propto \rho^{1/2-1}(1-\rho_j)^{1/2-1}(1+\varphi\rho_j)^{-1}$$

Alternative representation

Normal + Generalized Beta:

$$\beta_j \mid \rho_j \sim N(0, 1/\rho_j - 1)$$

$$p(\rho_j) \propto \rho^{1/2-1}(1-\rho_j)^{1/2-1}(1+\varphi\rho_j)^{-1}$$

Special Case: $\varphi = 1$ then

$$\rho_j \sim \text{Beta}(1/2, 1/2)$$

Alternative representation

Normal + Generalized Beta:

$$\beta_j \mid \rho_j \sim N(0, 1/\rho_j - 1)$$

$$p(\rho_j) \propto \rho^{1/2-1}(1-\rho_j)^{1/2-1}(1+\varphi\rho_j)^{-1}$$

Special Case: $\varphi = 1$ then

$$\rho_j \sim \text{Beta}(1/2, 1/2)$$

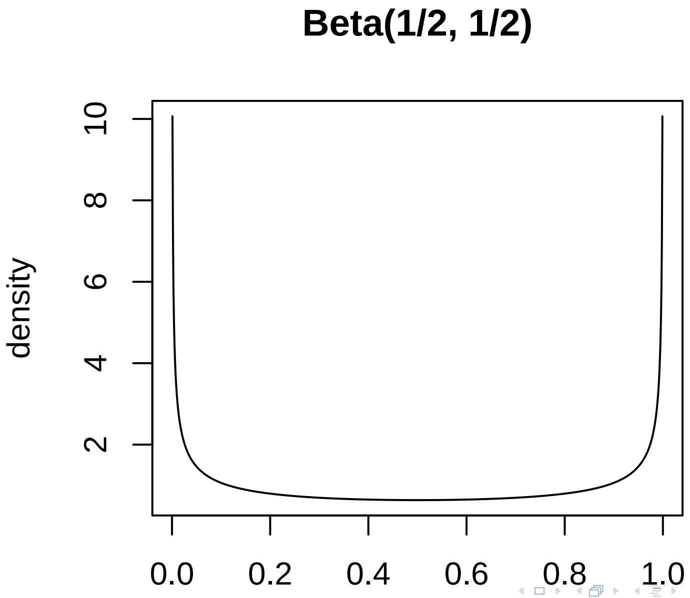
Induced Shrinkage:

$$\hat{\beta}_j \mid \beta_j \sim N(\beta_j, 1)$$

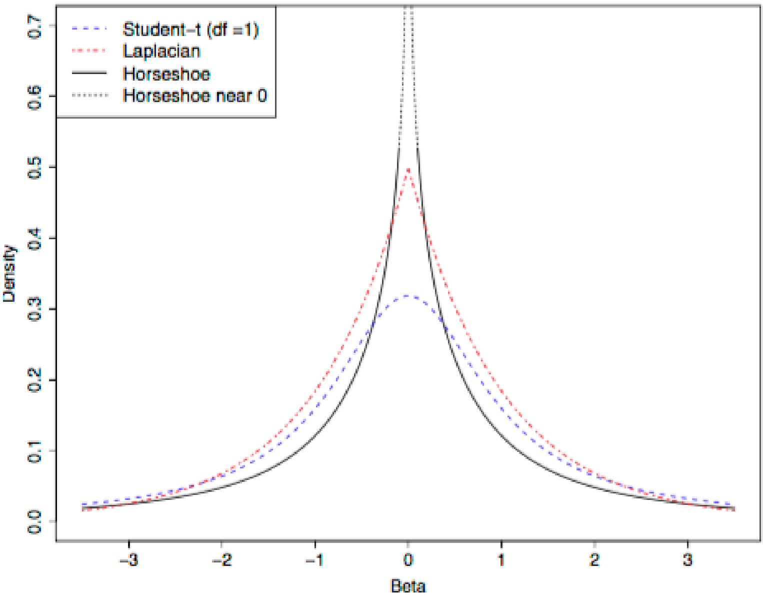
$$\beta_j \mid \mathbf{Y}, \rho_j \sim \mathrm{N}\left((1-\rho_j)\hat{\beta}_j, 1-\rho_j\right)$$

duke.eps

Horseshoe Prior Shrinkage



Prior



duke.eps

Outliers

Why should we assume that errors are normally distributed?

Outliers

Why should we assume that errors are normally distributed?

Use heavy tailed distributions for errors too! Student t, etc

Robust Regression with t errors

Model

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \omega_j, \phi \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1}\omega^{-1})$$

Robust Regression with t errors

Model

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \omega_j, \phi \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1}\omega^{-1})$$

$$\omega_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(\nu/2, \nu/2)$$

Robust Regression with t errors

Model

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \omega_j, \phi \stackrel{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1}\omega^{-1})$$

$$\omega_i \stackrel{\text{iid}}{\sim} \mathsf{Gamma}(\nu/2, \nu/2)$$

implies that marginally

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \phi \stackrel{\text{ind}}{\sim} \mathsf{St}(\nu, \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1/2})$$

Robust Regression with t errors

Model

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \omega_j, \phi \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1}\omega^{-1})$$

$$\omega_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(\nu/2, \nu/2)$$

implies that marginally

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \phi \overset{\text{ind}}{\sim} \mathsf{St}(\nu, \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1/2})$$

Interpretation of ω as latent weights

$$p(y_i) = (2\pi)^{-1/2}(\phi\omega_i)^{1/2} \exp(-\frac{\phi\omega_j}{2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2)$$

Robust Regression with t errors

Model

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \omega_j, \phi \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1}\omega^{-1})$$

$$\omega_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(\nu/2, \nu/2)$$

implies that marginally

$$Y_i \mid \boldsymbol{\beta}, \beta_0, \phi \overset{\text{ind}}{\sim} \mathsf{St}(\nu, \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}, \phi^{-1/2})$$

Interpretation of ω as latent weights

$$p(y_i) = (2\pi)^{-1/2}(\phi\omega_i)^{1/2} \exp(-\frac{\phi\omega_j}{2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2)$$

Small ω down weights errors

Conditional Distribution

Prior $\times$ Likelihood for case i

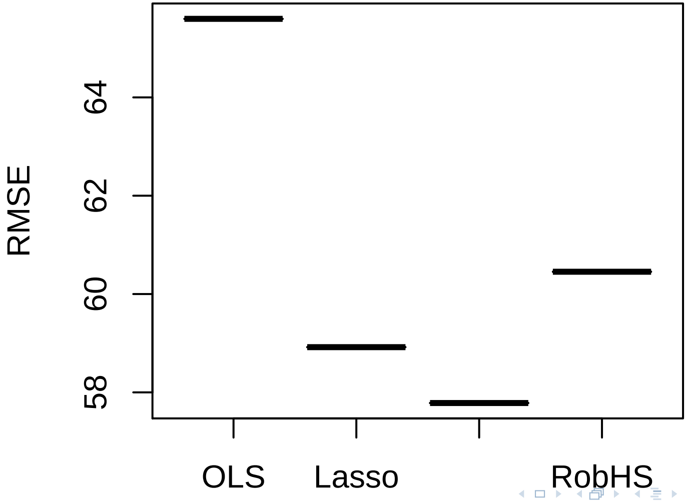
$$p(\omega_i \mid \cdot) \propto \omega_i^{\nu/2-1} \exp(-\frac{\nu}{2}\omega_i)(2\pi)^{-1/2}(\phi\omega_i)^{1/2} \exp\left(-\frac{\phi\omega_j}{2}(y_i - \mathbf{x}_i^T\beta)^2\right)$$

Code

```
# library(lars)
# library(monomvn)
# data(diabetes)

# yf = diabetes£y
# Xf = diabetes£x2
# do not center/scale as doen within blasso
# rbhs = blasso(Xf, yf, case="hs",
#                 theta = 16, RJ=FALSE,
#                 thin=10, T=2000,
#                 verb=0)
# y.pred = mean(rbhs£mu) +
#           Xf %*% apply(rbhs£beta, 2, mean)
```

Simulation Study with Diabetes Data



duke.eps

Other Options

Range of other scale mixtures used

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)
- Normal-Exponential-Gamma (Griffen & Brown 2005)

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)
- Normal-Exponential-Gamma (Griffen & Brown 2005)
- Relevance Vector Machines (Tipping) (improper!)

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)
- Normal-Exponential-Gamma (Griffen & Brown 2005)
- Relevance Vector Machines (Tipping) (improper!)
- Bridge - Power Exponential Priors (Stable mixing density)

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)
- Normal-Exponential-Gamma (Griffen & Brown 2005)
- Relevance Vector Machines (Tipping) (improper!)
- Bridge - Power Exponential Priors (Stable mixing density)

Some implemented in monomvn, but easy to add in JAGS

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)
- Normal-Exponential-Gamma (Griffen & Brown 2005)
- Relevance Vector Machines (Tipping) (improper!)
- Bridge - Power Exponential Priors (Stable mixing density)

Some implemented in monomvn, but easy to add in JAGS

Prior on β should have heavier tails than error distribution

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Bayesian Posterior under Shrinkage Priors does not assign any probability to $\beta_j = 0$

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Bayesian Posterior under Shrinkage Priors does not assign any probability to $\beta_j = 0$

- Selection solved as a post-analysis decision problem

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Bayesian Posterior under Shrinkage Priors does not assign any probability to $\beta_j = 0$

- Selection solved as a post-analysis decision problem
- Selection part of model uncertainty $\Rightarrow$ add prior

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Bayesian Posterior under Shrinkage Priors does not assign any probability to $\beta_j = 0$

- Selection solved as a post-analysis decision problem
- Selection part of model uncertainty $\Rightarrow$ add prior probability that $\beta_j = 0$ and combine with decision problem

Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection)

Bayesian Posterior under Shrinkage Priors does not assign any probability to $\beta_j = 0$

- Selection solved as a post-analysis decision problem
- Selection part of model uncertainty $\Rightarrow$ add prior probability that $\beta_j = 0$ and combine with decision problem
- Use 'RJ=TRUE' in blasso