Shrinkage Priors and Selection
Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 24, 2017
Bayesian Shrinkage

\[ Y \mid \alpha, \beta^s, \phi \sim N(1_n\alpha + X^s\beta^s, I_n/\phi) \]
Bayesian Shrinkage

\[ Y \mid \alpha, \beta^s, \phi \sim \text{N}(1_n \alpha + X^s \beta^s, I_n/\phi) \]
\[ \beta^s \mid \alpha, \phi, \tau, \lambda \sim \text{N}(0, \text{diag}(\tau^2)/\phi) \]
Bayesian Shrinkage

\[
Y \mid \alpha, \beta^s, \phi \sim N(1_n \alpha + X^s \beta^s, I_n / \phi)
\]
\[
\beta^s \mid \alpha, \phi, \tau, \lambda \sim N(0, \text{diag}(\tau^2) / \phi)
\]
\[
p(\alpha, \phi) \propto 1 / \phi
\]
Bayesian Shrinkage

\[
Y \mid \alpha, \beta^s, \phi \sim N(1_n \alpha + X^s \beta^s, I_n / \phi)
\]
\[
\beta^s \mid \alpha, \phi, \tau, \lambda \sim N(0, \text{diag}(\tau^2) / \phi)
\]
\[
p(\alpha, \phi) \propto 1 / \phi
\]

prior on \(\tau_j\)
Bayesian Shrinkage

\[ \mathbf{Y} \mid \alpha, \beta^s, \phi \sim \mathcal{N}(\mathbf{1}_n \alpha + \mathbf{X}^s \beta^s, \mathbf{I}_n/\phi) \]

\[ \beta^s \mid \alpha, \phi, \tau, \lambda \sim \mathcal{N}(\mathbf{0}, \text{diag}(\tau^2)/\phi) \]

\[ p(\alpha, \phi) \propto 1/\phi \]

prior on \( \tau_j \)

Scale Mixture of Normals (Andrews and Mallows 1974)
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

\[ \beta_s \mid \phi, \tau \sim N(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

\[ \beta^s \mid \phi, \tau \sim N(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \( \tau_j \mid \lambda \overset{\text{iid}}{\sim} C^+(0, \lambda^2) \) (difference in CPS notation)
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

\[ \beta^s \mid \phi, \tau \sim N(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \( \tau_j \mid \lambda \sim \text{C}^+(0, \lambda^2) \) (difference in CPS notation)
- \( \lambda \sim \text{C}^+(0, 1) \)
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on
  \[ \beta^s \mid \phi, \tau \sim N(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \[ \tau_j \mid \lambda \overset{\text{iid}}{\sim} C^+(0, \lambda^2) \] (difference in CPS notation)
- \[ \lambda \sim C^+(0, 1) \]
- \[ p(\alpha, \phi) \propto 1/\phi \]
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

\[ \beta^s \mid \phi, \tau \sim N(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \[ \tau_j \mid \lambda \overset{iid}{\sim} C^+(0, \lambda^2) \] (difference in CPS notation)
- \[ \lambda \sim C^+(0, 1) \]
- \[ p(\alpha, \phi) \propto 1/\phi \]

In the case \( \lambda = \phi = 1 \) and with canonical representation

\[ Y = \mathbf{I}_p \beta + \epsilon \]
Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

\[ \beta^s \mid \phi, \tau \sim \mathcal{N}(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \( \tau_j \mid \lambda \sim \mathcal{C}^+(0, \lambda^2) \) (difference in CPS notation)
- \( \lambda \sim \mathcal{C}^+(0, 1) \)
- \( p(\alpha, \phi) \propto 1/\phi \)

In the case \( \lambda = \phi = 1 \) and with canonical representation

\[ \mathbf{Y} = \mathbf{I} \beta + \epsilon \]

\[
E[\beta_i \mid \mathbf{Y}] = \int_0^1 (1 - \kappa_i)y_i^* p(\kappa_i \mid \mathbf{Y}) \ d\kappa_i = (1 - E[\kappa \mid y_i^*])y_i^*
\]

where \( \kappa_i = 1/(1 + \tau_i^2) \) shrinkage factor
Horseshoe

Carvalho, Polson & Scott propose

Prior Distribution on

\[ \beta^s | \phi, \tau \sim \mathcal{N}(0_p, \frac{\text{diag}(\tau^2)}{\phi}) \]

- \( \tau_j | \lambda \overset{\text{iid}}{\sim} \mathcal{C}^+(0, \lambda^2) \) (difference in CPS notation)
- \( \lambda \sim \mathcal{C}^+(0, 1) \)
- \( p(\alpha, \phi) \propto 1/\phi \)

In the case \( \lambda = \phi = 1 \) and with canonical representation \( \mathbf{Y} = \mathbf{I} \mathbf{\beta} + \epsilon \)

\[ E[\beta_i | \mathbf{Y}] = \int_0^1 (1 - \kappa_i) y_i^* p(\kappa_i | \mathbf{Y}) d\kappa_i = (1 - E[\kappa | y_i^*]) y_i^* \]

where \( \kappa_i = 1/(1 + \tau_i^2) \) shrinkage factor

Half-Cauchy prior induces a Beta(1/2, 1/2) distribution on \( \kappa_i \) a priori
Horseshoe

Beta(1/2, 1/2)
Prior Comparison (from PSC)
Bounded Influence

Normal means case $Y_i \overset{\text{iid}}{\sim} N(\beta_i, 1)$ (Equivalent to Canonical case)

- Posterior mean
  $$E[\beta \mid y] = y + \frac{d}{dy} \log m(y)$$
  where $m(y)$ is the predictive density under the prior (known $\lambda$)
Bounded Influence

Normal means case $Y_i \overset{iid}{\sim} N(\beta_i, 1)$ (Equivalent to Canonical case)

- Posterior mean
  $$E[\beta | y] = y + \frac{d}{dy} \log m(y)$$
  where $m(y)$ is the predictive density under the prior (known $\lambda$)

- HS has Bounded Influence:
  $$\lim_{|y| \to \infty} \frac{d}{dy} \log m(y) = 0$$
Bounded Influence

Normal means case $Y_i \sim N(\beta_i, 1)$ (Equivalent to Canonical case)

- Posterior mean
  
  \[ E[\beta | y] = y + \frac{d}{dy} \log m(y) \]

  where $m(y)$ is the predictive density under the prior (known $\lambda$)

- HS has Bounded Influence:

  \[
  \lim_{|y| \to \infty} \frac{d}{dy} \log m(y) = 0
  \]

- \[
  \lim_{|y| \to \infty} E[\beta | y) \to y
  \]

  (MLE)
Bounded Influence

Normal means case \( Y_i \overset{iid}{\sim} N(\beta_i, 1) \) (Equivalent to Canonical case)

- Posterior mean
  \[
  E[\beta \mid y] = y + \frac{d}{dy} \log m(y)
  \]
  where \( m(y) \) is the predictive density under the prior (known \( \lambda \))

- HS has Bounded Influence:
  \[
  \lim_{|y| \to \infty} \frac{d}{dy} \log m(y) = 0
  \]

- \( \lim_{|y| \to \infty} E[\beta \mid y] \to y \) (MLE)

- DE is also bounded influence, but bound does not decay to zero in tails
The `monomvn` package in R includes

- `lasso`
- `bhs`

See `Diabetes.R` code
Other Options

Range of other scale mixtures used

▶ Generalized Double Pareto (Armagan, Dunson & Lee)
  \[\tau_j \sim \text{Exp}\left(\frac{\lambda^2}{2}\right)\]
  \[\lambda \sim \text{Gamma}(\alpha, \eta)\]
  \[\beta_{sj} \sim \text{GDP}(\xi = \eta/\alpha, \alpha)\]
  \[f(\beta_{sj}) = \frac{1}{2^\xi (1 + |\beta_{sj}|^{\frac{1}{\alpha}})^{1 + \frac{1}{\alpha}}}\]

▶ Normal-Exponential-Gamma (Griffen & Brown 2005)
  \[\lambda^2 \sim \text{Gamma}(\alpha, \eta)\]

▶ Bridge - Power Exponential Priors (Stable mixing density)
  See the monomvn package on CRAN

Choice of prior? Properties?
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[
\begin{align*}
\tau_j & \sim \text{Exp}(\frac{\lambda}{2}) \\
\lambda & \sim \text{Gamma}(\alpha, \eta) \\
\beta_s j & \sim \text{GDP}(\xi = \frac{\eta}{\alpha}, \alpha) \\
f(\beta_s j) &= \frac{2}{\xi} (1 + \frac{\beta_s j}{\xi \alpha})^{-\frac{1+\alpha}{\alpha}}.
\end{align*}
\]


- Normal-Exponentale-Gamma (Griffin & Brown 2005)

- Bridge - Power Exponential Priors (Stable mixing density)

See the monomvn package on CRAN

Choice of prior? Properties?
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[ \tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2) \]

\[ \lambda \sim \text{Gamma}(\alpha, \eta) \]

\[ \beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha) \]


- Normal-Exponenetial-Gamma (Griffen & Brown 2005)

\[ \lambda^2 \sim \text{Gamma}(\alpha, \eta) \]

- Bridge - Power Exponential Priors (Stable mixing density)

See the monomvn package on CRAN

Choice of prior? Properties?
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[ \tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2) \]
\[ \lambda \sim \text{Gamma}(\alpha, \eta) \]
\[ \beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha) \]

\[ f(\beta_j^s) = \frac{1}{2\xi}(1 + \frac{|\beta_j^s|}{\xi\alpha})^{-(1+\alpha)} \]

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[
\tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2) \\
\lambda \sim \text{Gamma}(\alpha, \eta) \\
\beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha)
\]

\[
f(\beta_j^s) = \frac{1}{2\xi}(1 + \frac{|\beta_j^s|}{\xi\alpha})^{-(1+\alpha)}
\]


- Normal-Exponential-Gamma (Griffen & Brown 2005)

\[
\lambda^2 \sim \text{Gamma}(\alpha, \eta)
\]
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[ \tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2) \]
\[ \lambda \sim \text{Gamma}(\alpha, \eta) \]
\[ \beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha) \]

\[ f(\beta_j^s) = \frac{1}{2\xi} \left( 1 + \frac{|\beta_j^s|}{\xi \alpha} \right)^{-(1+\alpha)} \]


- Normal-Exponential-Gamma (Griffen & Brown 2005)

\[ \lambda^2 \sim \text{Gamma}(\alpha, \eta) \]

- Bridge - Power Exponential Priors (Stable mixing density)
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[ \tau_j^2 | \lambda \sim \text{Exp}(\lambda^2/2) \]
\[ \lambda \sim \text{Gamma}(\alpha, \eta) \]
\[ \beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha) \]

\[ f(\beta_j^s) = \frac{1}{2\xi}(1 + \frac{|\beta_j^s|}{\xi\alpha})^{-(1+\alpha)} \]


- Normal-Exponential-Gamma (Griffen & Brown 2005)

\[ \lambda^2 \sim \text{Gamma}(\alpha, \eta) \]

- Bridge - Power Exponential Priors (Stable mixing density)

See the monomvn package on CRAN
Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

\[ \tau_j^2 | \lambda \sim \text{Exp}(\lambda^2/2) \]
\[ \lambda \sim \text{Gamma}(\alpha, \eta) \]
\[ \beta_j^s \sim \text{GDP}(\xi = \eta/\alpha, \alpha) \]

\[ f(\beta_j^s) = \frac{1}{2\xi} \left( 1 + \frac{|\beta_j^s|}{\xi \alpha} \right)^{-(1+\alpha)} \]


- Normal-Exponential-Gamma (Griffen & Brown 2005)

\[ \lambda^2 \sim \text{Gamma}(\alpha, \eta) \]

- Bridge - Power Exponential Priors (Stable mixing density)

See the monomvn package on CRAN

Choice of prior? Properties?
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

\[
Y = X\beta + \epsilon
\]

Assume \(X^TX = I_p\) (orthonormal) and \(\epsilon \sim N(0, I_n)\)

Penalized Likelihood

\[
\frac{1}{2} \|Y - \hat{Y}\|^2_2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(\|\beta_j\|)
\]

Duality \(p_\lambda(\|\beta_j\|)\) is negative log prior

Requirements on penalty

- **Unbiasedness:** for large \(|\beta_j|\)
- **Sparsity:** thresholding rule sets small coefficients to 0
- **Continuity:** continuous in \(\hat{\beta}_j\)
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$

Penalized Likelihood

$$\frac{1}{2} \| Y - \hat{Y} \|^2_2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p(\lambda |\beta_j|)$$

Duality $p(\lambda |\beta_j|)$ is negative log prior

Requirements on penalty

- Unbiasedness: for large $|\beta_j|$%
- Sparsity: thresholding rule sets small coefficients to 0
- Continuity: continuous in $\hat{\beta}_j$
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$
- Assume $X^TX = I_p$ (orthonormal) and $\epsilon \sim N(0, I_n)$
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$
- Assume $X^T X = I_p$ (orthonormal) and $\epsilon \sim N(0, I_n)$
- Penalized Likelihood

$$\frac{1}{2} \|Y - \hat{Y}\|^2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(\|\beta_j\|)$$
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model \( Y = X\beta + \epsilon \)
- Assume \( X^TX = I_p \) (orthonormal) and \( \epsilon \sim N(0, I_n) \)
- Penalized Likelihood

\[
\frac{1}{2}\|Y - \hat{Y}\|^2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(\|\beta_j\|)
\]

duality \( p_\lambda(\|\beta\|) \) is negative log prior
- Requirements on penalty
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$
- Assume $X^TX = I_p$ (orthonormal) and $\epsilon \sim N(0, I_n)$
- Penalized Likelihood
  \[
  \frac{1}{2} \| Y - \hat{Y} \|^2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(|\beta_j|)
  \]
  duality $p_\lambda(\|\beta\|)$ is negative log prior
- Requirements on penalty
  - Unbiasedness: for large $|\beta_j|$
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$
- Assume $X^TX = I_p$ (orthonormal) and $\epsilon \sim N(0, I_n)$
- Penalized Likelihood

$$\frac{1}{2}\|Y - \hat{Y}\|^2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(\|\beta_j\|)$$

- duality $p_\lambda(\|\beta\|)$ is negative log prior
- Requirements on penalty
  - Unbiasedness: for large $|\beta_j|$
  - Sparsity: thresholding rule sets small coefficients to 0
Properties for Penalty for Modal Estimates

Fan & Li (JASA 2001) discuss Variable selection via nonconcave penalties and oracle properties

- Model $Y = X\beta + \epsilon$
- Assume $X^TX = I_p$ (orthonormal) and $\epsilon \sim N(0, I_n)$
- Penalized Likelihood

$$\frac{1}{2}\|Y - \hat{Y}\|^2 + \frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(|\beta_j|)$$

duality $p_\lambda(|\beta|)$ is negative log prior

- Requirements on penalty
  - Unbiasedness: for large $|\beta_j|
  - Sparsity: thresholding rule sets small coefficients to 0
  - Continuity: continuous in $\hat{\beta}_j$
Derivative of $\frac{1}{2} \sum_j (\beta_j - \hat{\beta}_j)^2 + \sum_j p_\lambda(|\beta_j|)$ is

$$\text{sgn}(\beta_j) \left\{ |\beta_j| + p'_\lambda(|\beta_j|) \right\} - \hat{\beta}_j$$

Conditions:

- unbiased: if $p'_\lambda(|\beta|) = 0$ for large $|\beta|$; estimator is $\hat{\beta}_j$
- thresholding: min $\{|\beta_j| + p'_\lambda(|\beta_j|)\} > 0$ then estimator is 0 if $|\hat{\beta}_j| < \min \{|\beta_j| + p'_\lambda(|\beta_j|)\}$
- continuity: minimum of $|\beta_j| + p'_\lambda(|\beta_j|)$ is at zero
Choice?

- Lasso does not satisfy conditions
- GDP does
Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)

Bayesian Posterior 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

Remember all models are wrong, but some may be useful!
Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection, just shrinkage) (Squared error loss)
- Minimize $L_1$ posterior loss $E[|\beta_j - a|]$ (Shrinkage and Selection)

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

Selection solved as a post-analysis decision problem

Selection part of model uncertainty ⇒ add prior probability that $\beta_j = 0$ and combine with decision problem

Remember all models are wrong, but some may be useful!
Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection, just shrinkage) (Squared error loss)
- Minimize $L_1$ posterior loss $E[|\beta_j - a|]$ (Shrinkage and Selection)

Bayesian Posterior does not assign any probability to $\beta_j^s = 0$
Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection, just shrinkage) (Squared error loss)
- Minimize $L_1$ posterior loss $E[|\beta_j - a|]$ (Shrinkage and Selection)

Bayesian Posterior does not assign any probability to $\beta_j^s = 0$

- Selection solved as a post-analysis decision problem
Choice of Estimator & Selection?

- Posterior Mode (may set some coefficients to zero)
- Posterior Mean (no selection, just shrinkage) (Squared error loss)
- Minimize $L_1$ posterior loss $E[|\beta_j - a|]$ (Shrinkage and Selection)

Bayesian Posterior 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, just shrinkage) (Squared error loss)
- Minimize $L_1$ posterior loss $E[|\beta_j - a|]$ (Shrinkage and Selection)

Bayesian Posterior does not assign any probability to $\beta_j^s = 0$

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

Remember all models are wrong, but some may be useful!