Shrinkage Priors and Selection

Readings Chapter 15 Christensen

STA721 Linear Models Duke University

Merlise Clyde

October 28, 2019

Bayesian Shrinkage

$$\mathbf{Y} \mid \alpha, \boldsymbol{\beta}^s, \phi \sim \mathsf{N}(\mathbf{1}_n\alpha + \mathbf{X}^s\boldsymbol{\beta}^s, \mathbf{I}_n/\phi)$$

Bayesian Shrinkage

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

Bayesian Shrinkage

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

Bayesian Shrinkage

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

prior on τ_j

Bayesian Shrinkage

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

prior on τ_j
Scale Mixture of Normals (Andrews and Mallows 1974)

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}_p, \frac{\mathsf{diag}(\boldsymbol{\tau}^2)}{\phi})$$

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}_p, \frac{\mathsf{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}_p, \frac{\mathsf{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}_p, \frac{\text{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathrm{N}(\mathbf{0}_p, \frac{\mathrm{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

In the case $\lambda = \phi = 1$ and with canonical representation $\mathbf{Y}^* = \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathrm{N}(\mathbf{0}_p, \frac{\mathrm{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

In the case $\lambda = \phi = 1$ and with canonical representation $\mathbf{Y}^* = \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

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

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

Horseshoe

Carvalho, Polson & Scott propose

- Prior Distribution on

$$\boldsymbol{\beta}^s \mid \phi, \boldsymbol{\tau} \sim \mathsf{N}(\mathbf{0}_p, \frac{\text{diag}(\boldsymbol{\tau}^2)}{\phi})$$

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

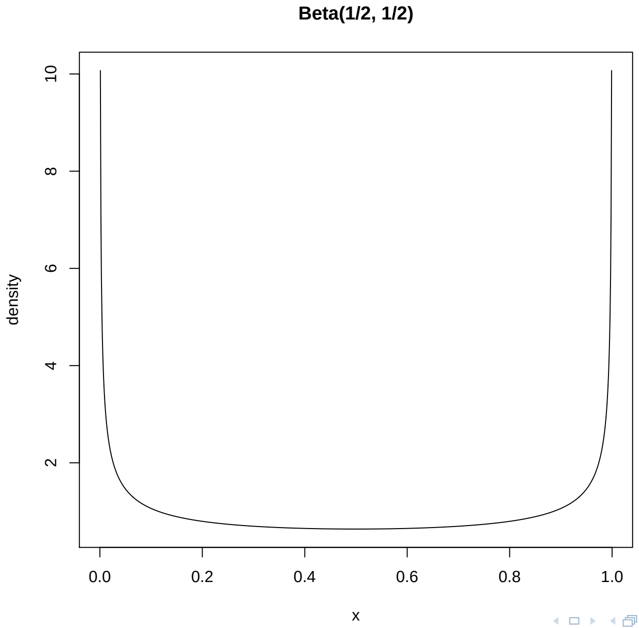
In the case $\lambda = \phi = 1$ and with canonical representation $\mathbf{Y}^* = \mathbf{I}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

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

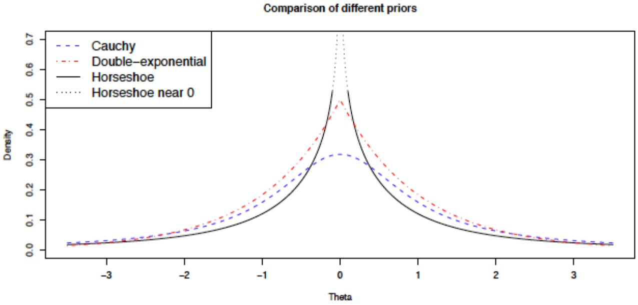
where $\kappa_i = 1/(1 + \tau_i^2)$ shrinkage factor

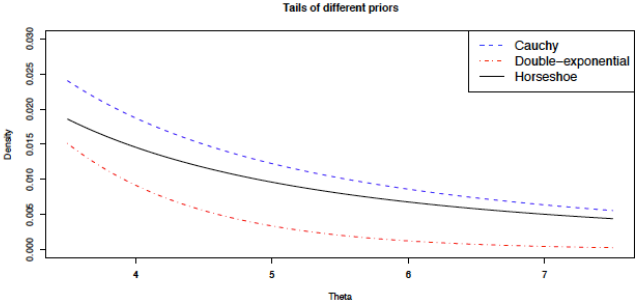
Half-Cauchy prior induces a Beta(1/2, 1/2) distribution on κ_i a priori

Horseshoe



Prior Comparison (from PSC)





Bounded Influence

Normal means case $Y_i \stackrel{\text{iid}}{\sim} N(\beta_i, 1)$ (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 λ)

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 λ)
- HS has Bounded Influence:

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

Bounded Influence

Normal means case $Y_i \stackrel{\text{iid}}{\sim} N(\beta_i, 1)$ (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 λ)
- 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)

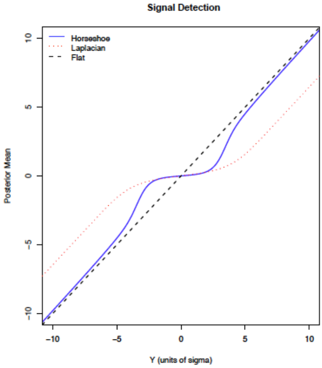
Bounded Influence

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 λ)
- 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



R packages

The `monomvn` package in R includes

- blasso
- bhs

See Diabetes.R code

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)

$$
\begin{aligned}
\tau_j^2 \mid \lambda &\sim \mathrm{Exp}(\lambda^2/2) \\
\lambda &\sim \mathrm{Gamma}(\alpha, \eta) \\
\beta_j^s &\sim \mathrm{GDP}(\xi = \eta/\alpha, \alpha)
\end{aligned}
$$

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

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

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

see `http://arxiv.org/pdf/1104.0861.pdf`

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

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

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

  see `http://arxiv.org/pdf/1104.0861.pdf`

- Normal-Exponenetial-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)

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

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

  see `http://arxiv.org/pdf/1104.0861.pdf`

- Normal-Exponenetial-Gamma (Griffen & Brown 2005)
  $\lambda^2 \sim \mathsf{Gamma}(\alpha, \eta)$
- Bridge - Power Exponential Priors (Stable mixing density)

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

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

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

  see http://arxiv.org/pdf/1104.0861.pdf

- 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

Other Options

Range of other scale mixtures used

- Generalized Double Pareto (Armagan, Dunson & Lee)

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

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

  see `http://arxiv.org/pdf/1104.0861.pdf`

- Normal-Exponenetial-Gamma (Griffen & Brown 2005)
  $\lambda^2 \sim \mathsf{Gamma}(\alpha, \eta)$
- 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

Properties for Penalty for Modal Estimates

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

- Model $Y = \mathbf{X}\beta + \epsilon$

Properties for Penalty for Modal Estimates

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

- Model $Y = \mathbf{X}\beta + \epsilon$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\epsilon \sim N(0, \mathbf{I}_n)$

Properties for Penalty for Modal Estimates

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

- Model $Y = \mathbf{X}\beta + \epsilon$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\epsilon \sim N(0, \mathbf{I}_n)$
- Penalized Likelihood

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

Properties for Penalty for Modal Estimates

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

- Model $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n)$
- Penalized Likelihood

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

  duality $p_\lambda(|\beta|)$ is negative log prior
- Requirements on penality

Properties for Penalty for Modal Estimates

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

- Model $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n)$
- Penalized Likelihood

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

  duality $p_\lambda(|\beta|)$ is negative log prior
- Requirements on penality
  - 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 $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n)$
- Penalized Likelihood

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

  duality $p_\lambda(|\beta|)$ is negative log prior
- Requirements on penality
  - 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 $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$
- Assume $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ (orthonormal) and $\epsilon \sim N(0, \mathbf{I}_n)$
- Penalized Likelihood

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

  duality $p_\lambda(|\beta|)$ is negative log prior
- Requirements on penality
  - Unbiasedness: for large $|\beta_j|$
  - Sparsity: thresholding rule sets small coefficients to 0
  - Continuity: continuous in $\hat{\beta}_j$

Conditions

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

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

Conditions:

- unbiased: if $p'_\lambda(|\beta|) = 0$ for large $|\beta|$; estimator is $\hat{\beta}_j$
- thresholding: $\min\left\{|\beta_j| + p'_\lambda(|\beta_j|)\right\} > 0$ then estimator is 0 if $|\hat{\beta}_j| < \min\left\{|\beta_j| + p'_\lambda(|\beta_j|)\right\}$
- 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)

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 $\mathrm{E}[|\beta_j - a|]$ (Shrinkage and Selection)

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 $\mathrm{E}[|\beta_j - a|]$ (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 $\mathrm{E}[|\beta_j - a|]$ (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 $\mathrm{E}[|\beta_j - a|]$ (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

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 $\mathrm{E}[|\beta_j - a|]$ (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!