The Bayesian Lasso (Park & Casella; 2008)

Trevor Park & George Casella

Department of Statistics
University of Kentucky

Dec 04, 2019

Penalized Regression

- Penalized regression by solving (Frank and Friedman 1993)

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|^q,$$

  for some $q \geq 0$
- The Bayesian analog of this penalization involves using a prior on $\boldsymbol{\beta}$ of the form

$$\pi(\boldsymbol{\beta}) \propto \prod_j \exp(-\lambda |\beta_j|^q)$$

- Thus the elements of $\boldsymbol{\beta}$ have independent priors from the *exponential power distribution*.

Bimodality under the Unconditional Prior

- Put unconditional Laplace prior on $\boldsymbol{\beta}$ with some independent prior $\pi(\sigma^2)$ on σ^2

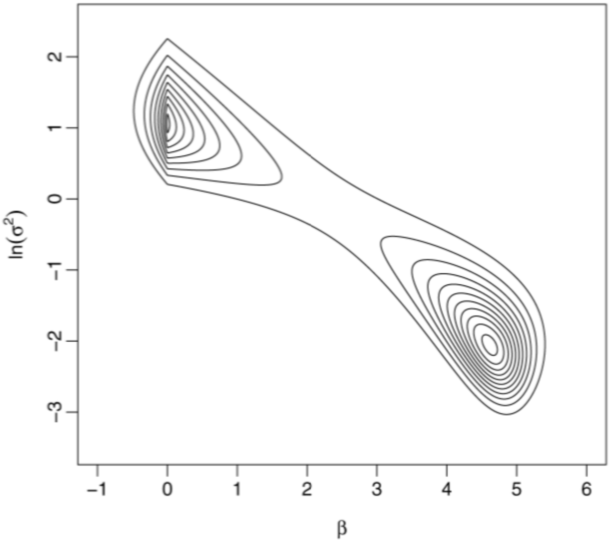
$$\prod_j \frac{\pi}{2} \exp(-\lambda|\beta_j|)\pi(\sigma^2)$$

- Then the joint posterior distribution

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto \pi(\sigma^2)(\sigma^2)^{-(n-1)/2} \times \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \lambda \sum_j |\beta_j|\}$$

- Posteriors of this form can easily have more than one mode.
- Example: $p = 1$, $n = 10$, $\mathbf{X}^T\mathbf{X} = 1$, $\mathbf{X}^T\mathbf{y} = 5$, $\mathbf{y}^T\mathbf{y} = 26$, $\lambda = 3$
  1. Least square: $\hat{\beta} = 5$, $\hat{\sigma}^2 = 1/8$
  2. Infinity penalty: $\hat{\beta} = 0$, $\hat{\sigma}^2 = 26/9$

Bimodality under the Unconditional Prior



Bimodality under the Unconditional Prior

Issue with Bimodality:

- Slows convergence of the Gibbs sampler.
- Point estimates less meaningful.

Solution: conditionally independent priors from the exponential power distribution

- $$\pi(\boldsymbol{\beta} \mid \sigma^2) \propto \prod_j \exp\{-\lambda(|\beta_j|/\sqrt{\sigma^2})^q\}$$
- Whenever $0 < q \leq 2$, this distribution may be represented by a scale mixture of normals:

$$\exp(-|z|^q) \propto \int_0^\infty \frac{1}{2\pi s} \exp(-\frac{z^2}{2s}) \frac{1}{s^{3/2}} g_{q/2}(\frac{1}{2s})\ ds,$$

where $g_{q/2}$ is a density of stable random variable with index $q/2$ which generally does not have close-form expression.

Unimodality under the Conditional Prior

- Put conditional Laplace prior on $\boldsymbol{\beta}$ with some independent prior $\pi(\sigma^2)$ on σ^2

$$\pi(\sigma^2)\prod_j \frac{\pi}{2\sqrt{\sigma^2}}\exp(-\lambda\frac{|\beta_j|}{\sqrt{\sigma^2}})$$

- Then the joint posterior distribution $\pi(\boldsymbol{\beta},\sigma^2 \mid \mathbf{y})$ is unimodal for typical choices of $\pi(\sigma^2)$ and any choice of $\lambda \geq 0$ in a sense that for every $c > 0$ the upper level set $\{(\boldsymbol{\beta},\sigma^2\mathbf{y}) : \pi(\boldsymbol{\beta},\sigma^2 \mid \mathbf{y}) \geq c, \sigma^2 > 0\}$ is connected.
- Dropping terms that involve neither $\boldsymbol{\beta}$ and σ^2, the log posterior is

$$\log(\pi(\sigma^2)) - \frac{n+p-1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 - \lambda\|\boldsymbol{\beta}\|_1/\sqrt{\sigma^2}$$

- Taking $\phi = \boldsymbol{\beta}/\sqrt{\sigma^2}, \rho = 1/\sqrt{\sigma^2}$

$$\log(\pi(1/\rho^2)) + \underbrace{(n+p-1)\log(\rho)}_{\text{concave}} - \underbrace{\frac{1}{2}\|\rho\mathbf{y}-\mathbf{X}\phi\|_2^2}_{\text{convex quadratic}} - \underbrace{\lambda\|\phi\|_1}_{\text{convex}}$$

- Hence the posterior is unimodal, provided that $\log(\pi(1/\rho^2))$ is concave.

Fully Bayes Lasso

Representation of the Laplace distribution

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{2\pi s}\exp(-\frac{z^2}{2s})\frac{a}{2}\exp(-a^2 s/2)\ ds$$

- Likelihood:

$$\mathbf{y} \mid \mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim N(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

- Prior:

$$\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N(\mathbf{0}, \sigma^2 \mathbf{D}_\tau),$$

  with $\mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$.
- Hyper parameter:

$$\sigma, \tau_1^2, \ldots, \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_j \frac{\lambda^2}{2}\exp(-\frac{\lambda^2 \tau_j^2}{2}) d\tau_j^2$$

- The parameter μ may be given an independent, flat prior.
- Park & Casella (2008) used improper prior density $\pi(\sigma^2) = 1/\sigma^2$, but any inverse-gamma prior for σ^2 also would maintain conjugacy.

Gibbs Sampler

- Because μ is rarely of interest, marginalize it out in the interest of simplicity and speed
- Marginalizing over μ does not affect conjugacy, which means full conditional distributions are still easy to sample
- Full conditional for β: $N((\mathbf{X}^T\mathbf{X}+\mathbf{D}_\tau^{-1})^{-1}\mathbf{X}\mathbf{y}, \sigma^2(\mathbf{X}^T\mathbf{X}+\mathbf{D}_\tau^{-1})^{-1})$
- Full conditional for σ^2: inverse gamma with shape $\dfrac{n-1}{2}+\dfrac{p}{2}$ and scale $(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)/2+\beta^T\mathbf{D}_\tau^{-1}\beta/2$
- Full conditional for $\tau_1^2,\ldots,\tau_p^2$: $\dfrac{1}{\tau_j^2}$ conditionally inverse Gaussian with parameter $\mu=\sqrt{\dfrac{\lambda^2\sigma^2}{\beta_j^2}}$ and $\lambda'=\lambda^2$, with

$$f(x)=\sqrt{\frac{\lambda'}{2\pi}}x^{-3/2}\exp\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2}x\}I_{(0,\infty)}(x)$$

Marginal Maximum Likelihood

- kth iteration of the algorithm involves running the Gibbs sampler using a $\lambda^{(k)}$ value estimated from the sample of the previous iteration, $\lambda^{(k-1)}$

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} \mathrm{E}_{\lambda^{(k-1)}}(\tau_j^2 \mid \mathbf{y})}}$$

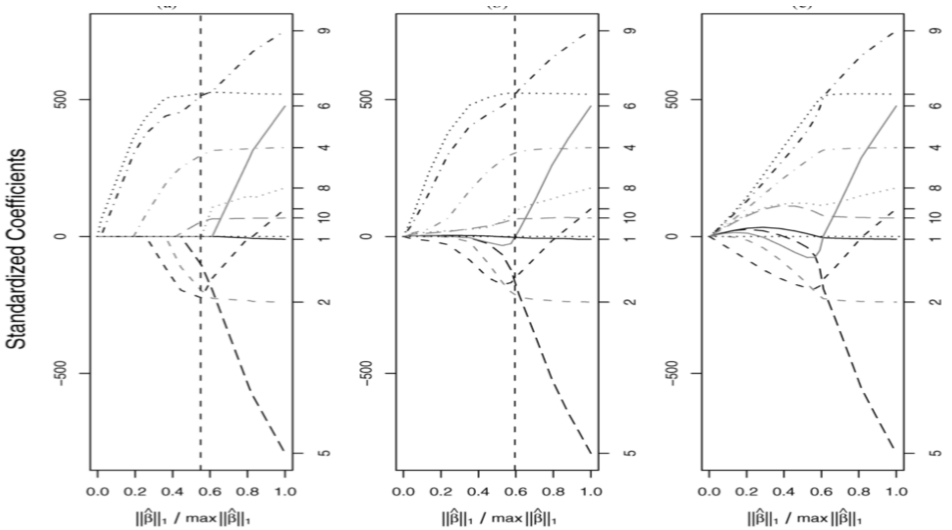
- The conditional expectation is replaced by the average from Gibbs sample.
- Initial value was suggested as $\lambda^{(0)} = \dfrac{p\sqrt{\hat{\sigma}_{\mathrm{LS}}}}{\sum_{j=1}^{p} |\hat{\beta}_j^{\mathrm{LS}}|}$

Hyperpriors

- The prior density for λ^2 (not λ) should approach 0 sufficiently fast as $\lambda^2 \to \infty$ (to avoid mixing problems) but should be relatively flat and place high probability near the maximum likelihood estimate.
- Gamma priors on λ^2 resulting conjugacy allows easy extension of the Gibbs sampler.

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{(r-1)} \exp(-\delta\lambda^2)$$

Comparing with Lasso and Ridge



Comparison on Real Data

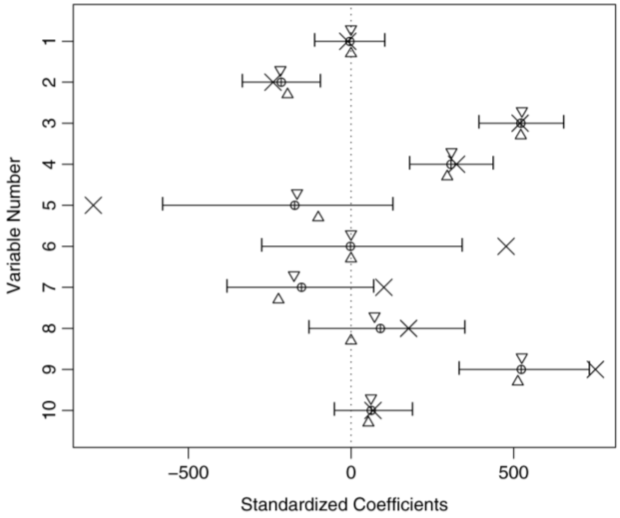


Figure 2. Posterior median Bayesian Lasso estimates (⊕) and corresponding 95% credible intervals (equal-tailed) with λ selected according to marginal maximum likelihood (Sec. 3.1). Overlaid are the least squares estimates (×), Lasso estimates based on n-fold cross-validation (△), and Lasso estimates chosen to match the L_1 norm of the Bayes estimates (▽). The variables were described by Efron et al. (2004): (1) age, (2) sex, (3) bmi, (4) map, (5) tc, (6) ldl, (7) hdl, (8) tch, (9) ltg, and (10) glu.