Lasso and Bayesian Lasso

Qi Tang

Department of Statistics
University of Wisconsin-Madison

Feb. 05, 2010

Outline

- Lasso (Tibshirani, 1996)

- The Bayesian Lasso (Park and Casella, 2008)

Variable Selection

- Why?
  - Interpretation: principle of parsimony.
  - Prediction: bias and variance tradeoff.
- What if number of variables is greater than number of observations ($p > n$)?
- Shrinkage!
  - loss function + penalty function. Ridge regression, Lasso (Tibshirani, 1996) and other methods.

Lasso (Tibshirani, 1996)

- Consider linear regression model

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n),$$

where y is the centered response ($\sum_{i=1}^{n} y_i = 0$); $X_1, \ldots, X_p$, columns of X, are centered to have 0 mean and standardized to have unit L_2 norm.

- The Lasso method solves the following optimization problem

$$\min_{\beta}\{\|y - X\beta\|^2\} \quad \text{subject to} \sum_{i=1}^{p} |\beta_i| \leq t \qquad (1)$$

where t needs to be tuned by cross validation.

Why Lasso can Set Some β_i to be 0?

- The loss function $\|y - X\beta\|^2$ equals to the quadratic function

$$(\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) + \text{constant}, \qquad (2)$$

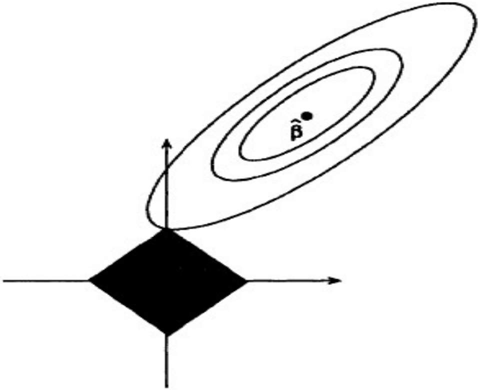
where $\hat{\beta}$ is the least square estimate.

- Consider the case $p = 2$.

- The constraint $|\beta_1| + |\beta_2| \leq t$ is a diamond region in the R^2 space.

Why Lasso can Set Some β_i to be 0?

- Curves are the contours of (2).
- The rotated square is the constraint region.
- Lasso solution is the place where the contour first touches the square.



Bayesian Interpretation of Lasso

- Lasso problem can be written into:

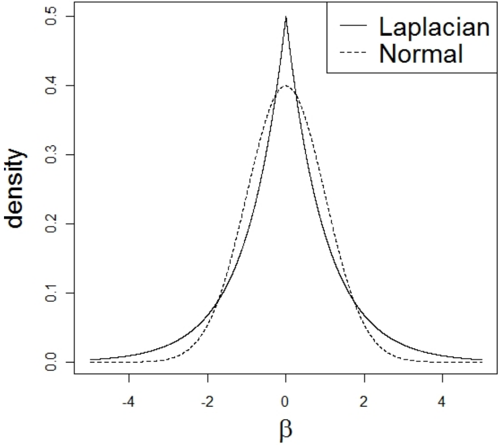
$$\min_{\beta}\{\|y - X\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|\} \tag{3}$$

- Consider the Bayesian model $y \sim N(X\beta, I_n)$ and $\beta_i \sim \frac{\lambda}{2} e^{-\lambda|\beta_i|}$ (Laplacian prior).

- The solution of (3) can be interpreted as the posterior mode of β in the above Bayesian model.

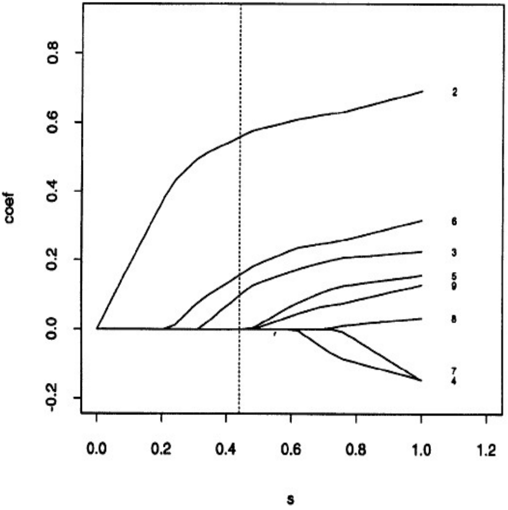
Laplacian Priors

- The Laplacian prior assigns more weight to regions near zero than the normal prior.



The Prostate Cancer Example

- $s = t / |\hat{\beta}|_{L_1}$.
- The broken line is at $s = 0.44$.



The Bayesian Lasso (Park and Casella, 2008)

- Model $y \mid X, \beta, \sigma^2 \sim N(X\beta, \sigma^2)$.

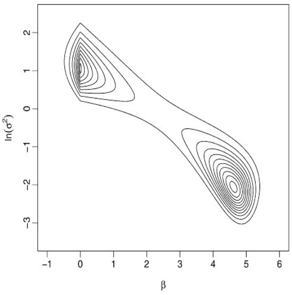
- Set the conditional Laplacian prior to β_i

$$\beta_i \mid \sigma^2 \sim \frac{\lambda}{2\sigma} e^{-\lambda|\beta_i|/\sigma},$$

  where conditioning on σ^2 is important to guarantee a unique posterior mode.

Unconditional Prior May Lead to Bimodal Posteriors

- Consider $\beta_i \sim \frac{\lambda}{2} e^{-\lambda|\beta_i|}$ with $p = 1$, $n = 10$, $X^T X = 1$, $X^T y = 5$, $y^T y = 26$ and $\lambda = 3$.
- The posterior distributions of $(\ln\sigma^2, \beta)$ are bimodal.



Rewrite the Laplacian Prior

- It can written into a mixture of the following hierarchical priors (integrating out γ_i^2)

$$\beta_i \mid (\sigma^2, \gamma_i^2) \sim N(0, \sigma^2\gamma_i^2) \quad \gamma_i^2 \mid \sigma^2 \sim Exp(\lambda^2/2). \tag{4}$$

- The reason is

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2} ds, \quad a > 0$$

Empirical Treatment of λ

- Estimate λ by the marginal maximum likelihood. Use the MCEM algorithm and update the value of λ by

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}[\gamma_i^2|y]}}.$$

- Assign a hyperprior to λ^2 that places high density at the marginal maximum likelihood estimate.

The Full Conditional Distributions

Assign $\pi(\sigma^2) = 1/\sigma^2$, then we have

- $\beta \sim N(A^{-1}X^Ty, \sigma^2 A^{-1})$, $A = X^TX + D_\gamma^{-1}$ and $D_\gamma = \text{diag}(\gamma_1^2, \ldots, \gamma_p^2)$.
- $\sigma^2 \sim \text{InvGammma}(a, b)$ with shape parameter $a = (n+p)/2$ and scale parameter $b = (y - X\beta)^T(y - X\beta)/2 + \beta^T D_\gamma^{-1}\beta/2$.
- $1/\gamma_i^2 \sim \mathit{InvGuassian}(a, b)$ with $a = \sqrt{\lambda^2\sigma^2/\beta_j^2}$ and $b = \lambda^2$.
- The inverse Guassian distribution with parameter a and b is of the following form:

$$f(x) = \frac{b}{2\pi}x^{-3/2} exp\left\{-\frac{b(x-a)^2}{2(a)^2x}\right\}, \quad x > 0.$$