Md Abul Hayat
University of Arkansas

Spring 2018
mahayat@uark.edu

Paper Review : STAT 5443 (Computational Statistics)

The Bayesian Lasso
Trevor Park and George Casella
Journal of the American Statistical Association
Volume 103, 2008 - Issue 482

Review:
Section 1 of this paper has discussed the basic concepts of Bayesian LASSO. A simple data-set has been used to illustrate the comparison between LASSO and Bayesian LASSO. Section 2 discusses a hierarchical model of a Bayesian LASSO and Section 3 implements the hierarchy using Gibbs sampling. Section 4 discusses about posterior distribution. Selection of hyper-parameters is discussed in Section 5. Section 6 discusses the extension of this idea.

LASSO is a popular sparse linear regression technique which is extensively used for feature selection. Suppose, we have a linear regression model

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, μ is the overall mean, $\boldsymbol{X}$ is the $n \times p$ matrix of standardised regressors, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the vector of regression coefficients to be estimated, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and unknown variance σ^2. The LASSO estimates can be found by solving the following convex optimization problem.

$$\min_{\beta} \quad (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_i| \tag{2}$$

where, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \mu \mathbf{1}_n$ is the centered response vector. The lasso solution is unique when $\text{rank}(X) = p$, because in this case the optimization criterion is strictly convex. This is not true when $\text{rank}(X) < p$, which happens when the number of variables exceeds the number of observations, i.e. $p > n$; in this case we must have $\text{rank}(X) < p$ and the LASSO fails to capture p features.

The LASSO problem can be alternatively approached using Bayesian methods. If the parameters β_i have independent and identical double exponential (Laplace) priors, then

$$\pi(\boldsymbol{\beta}) = \prod_{i=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_i|} = \frac{\lambda}{2} e^{-\lambda \sum_{i=1}^{p} |\beta_i|} \tag{3}$$

and from equation (1), we can think of $\tilde{\boldsymbol{y}}$ as

$$\tilde{\boldsymbol{y}} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

Likelihood,

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \tag{4}$$

Assuming, $\boldsymbol{\beta}$ and σ^2 to be independent, let σ^2 has prior distribution of $\pi(\sigma^2)$

$$p(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}}) = \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \sigma^2) \cdot \pi(\boldsymbol{\beta}) \cdot \pi(\sigma^2)}{p(\tilde{\boldsymbol{y}})} \propto \frac{\pi(\sigma^2)}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) - \lambda \sum_{i=1}^{p} |\beta_i| \right\} \tag{5}$$

from equation (5) and (2), we can say that for a fixed value of σ^2 and λ, the LASSO estimate which minimizes equation (2) is also the posterior mode estimate using (5). Both the problems in (2) and (5) are same except we have more flexibility on finding the posterior distribution of $\boldsymbol{\beta}$ using parameters like σ^2 and λ.

A more generalized and flexible hierarchical model of the same problem is as follows

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

$$\boldsymbol{\beta}|(\tau_1^2, \tau_2^2, \cdots, \tau_p^2, \sigma^2) \sim N_p(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_r), \quad \boldsymbol{D}_r = \text{diag}(\tau_1^2, \tau_2^2, \cdots, \tau_p^2) \tag{6}$$

$$\tau_1^2, \tau_2^2, \cdots, \tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2 \quad \tau_1^2, \tau_2^2, \cdots, \tau_p^2 > 0$$

$$\sigma^2 \sim \pi(\sigma^2) d\sigma^2$$

with $\tau_1^2, \tau_2^2, \cdots, \tau_p^2$ and σ^2 independent. Using $\sigma^2 \sim \text{IGamma}(a, \gamma)$ and $\pi(\mu) \propto 1$ a full conditional Gibbs sampler is implemented. The number of modes in posterior distribution depends on the proper selection of priors. If the posterior distribution has more than one modes then it is difficult to say if a single mode can describe the whole regression system. Computationally speaking, posterior distribution with more than one modes are hard to realize and needs care while implementing.

Methods like Expectation-Maximization can be used to estimate the value of λ. We can also use a prior distribution of λ, but care should be taken otherwise it may end-up with multi-modal distribution. Bridge regression or the generalized version of LASSO where L_q norm is used, where $q \in \mathbf{N}$ can also be implemented using the Bayesian LASSO concept using proper priors of $\boldsymbol{\beta}$.

In conclusion we can say that although the Bayesian LASSO does not automatically perform variable selection, it can provide a credible interval for each of the coefficients. If a parameter has to meet a certain threshold to be considered significant, a credible set will indicate the degree of certainty that this requirement is met.
