Sparse modifying algorithm in Bayesian lasso

Ibuki HOSHINA* and Sadanori KONISHI†

Abstract — The lasso is simultaneous variable selection and parameter estimation procedure in linear regression models. The estimates can be interpreted as a Bayesian posterior mode when independent Laplace prior distributions are placed on the regression coefficients. Park and Casella (2008) extended the Bayesian lasso linear regression model by placing prior distributions on hyper-parameters in independent Laplace distribution. It might be however noted that the point estimate of Bayesian lasso is not sparse. In the present paper, we propose an efficient algorithm which modifies the Bayesian lasso estimates so as to be sparse. Monte Carlo simulations are conducted to investigate the efficiency of the proposed algorithm.

Keyword: Bayesian lasso, Gibbs sampler, Regularization, Sparse regression, Variable selection

1 Introduction

The ordinary least square or maximum likelihood estimation often has poorly in both prediction accuracy and interpretation of linear regression models with a large number of predictors. In order to overcome this drawback, the lasso proposed by Tibshirani [9] has sparked much interest in the use of regularization for simultaneous variable selection and estimation. The lasso is a penalized least square method, imposing an $L_1$ norm penalty on regression coefficients. Due to the nature of the $L_1$ penalty, it shrinks some components of coefficients to zero. These nice properties make it a variable selection method. In regression analysis, the lasso has been extended by imposing different penalties for regression coefficients, and various $L_1$ regularization procedures have been proposed in literature (e.g., Zou and Hastie [10], Zou [11]).

Tibshirani [9] suggested that the lasso estimates can be interpreted as a Bayesian posterior mode when Laplace prior distributions are placed on the regression coefficients. Park and Casella [8] extended the Bayesian linear regression model by imposing prior distributions on hyperparameters in Laplace distribution, and used Gibbs sampling for parameter estimation (see also Hans [4]).

Although the lasso was originally designed for variable selection, the Bayesian lasso estimates proposed by Park and Casella [8] lose this attractive property, not setting any of the coefficients to zero. In order to overcome this drawback of non-sparsity, we propose an algorithm “Sparse modifying algorithm” that it creates a flexible numerical instrument for setting some coefficients identically equal to zero. The proposed algorithm is computationally-efficient, because there is no need to perform iterative calculation to obtain the solutions. Furthermore, our method can be applied to a wide variety of lasso-type regularized Bayes estimation such as the Bayesian elastic net (Li and Lin [7], Kyung et al. [6]). Monte Carlo simulations are conducted to investigate the efficiency of the proposed sparse modifying algorithm.

2 Bayesian lasso

We consider a linear regression model with $n$ observations and $p$ predictors:

$$y = \mu_1 + X\beta + \varepsilon$$

(2.1)

where $y = (y_1, \ldots, y_n)^T$ is the $n \times 1$ response vector, $\mu$ is the overall mean, $X$ is the $n \times p$ matrix of standardized regressors, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the coefficients, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the $n \times 1$ vector of independent and identically distributed errors with $N(0, \sigma^2)$. The lasso estimate of $\beta$ (Tibshirani [9]) is given as a solution of the $L_1$ penalized least squares:

$$\hat{\beta}^{lasso} = \arg \min_\beta ||y - X\beta||^2 + \lambda ||\beta||,$$

(2.2)

*Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan, Email: ibu@agc.math.chuo-u.ac.jp
†Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan, Email: konishi@math.chuo-u.ac.jp
where \( \tilde{y} = y - \hat{y}_1 \), \( |\beta| = \sum_{j=1}^{p} |\beta_j| \) is the \( L_1 \)-norm of \( \beta \), and \( \lambda \) is a positive tuning parameter. In (2.2), the \( L_1 \) penalty may set some components of \( \beta \) to zero for some appropriately chosen \( \lambda \), resulting in simultaneous estimation and variable selection procedure (We say the "sparsity" for this property). The entire lasso solution paths can be computed by the LARS algorithm (Efron et al. [2]) and the Coordinate Descent algorithm (Friedman et al. [3]).

Tibshirani [9] suggested that with the \( L_1 \) penalty term in (2.2), the lasso estimates could be interpreted as the Bayes posterior mode under independent Laplace priors

\[
\pi(\beta|\lambda) \propto \prod_{j=1}^{p} \exp\left( - |\beta_j| \lambda \right) \tag{2.3}
\]

for the regression coefficients. Park and Casella [8] extended it as a fully Bayesian model using a conditional Laplace prior specification of the form

\[
\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2 \sqrt{\sigma^2}} \exp\left( - \frac{\lambda |\beta_j|}{\sqrt{\sigma^2}} \right) \tag{2.4}
\]

The \( \sigma^2 \) has the noninformative scale-invariant marginal prior \( \pi(\sigma^2) = 1/\sigma^2 \) to avoid the multi-modality of full posterior distribution (They mentioned that any inverse-gamma prior for \( \sigma^2 \) also would maintain conjugacy).

It follows from Andrews and Mallows [1] that the Laplace density function can be expressed as

\[
a \frac{1}{2} \exp(-a|z|) = \int_0^{\infty} \frac{1}{\sqrt{2\pi s}} \exp\left( - \frac{z^2}{2s} \right) a \exp\left( \frac{a^2 s}{2} \right) ds. \tag{2.5}
\]

That is, the Laplace prior is equivalent to a two-level hierarchical Bayes model: scale mixture Gaussian distribution with independent exponentially distributed variances. The Bayesian lasso is based on this connection. Park and Casella [8] used the Gibbs sampling to obtain the Bayesian lasso estimates with the Laplace prior in the hierarchical model.

The full model of the Bayesian lasso is given in the following:

\[
\psi \mid \mu, X, \beta, \sigma^2 \sim N_n(\mu I_n + X\beta, \sigma^2 I_n), \quad \beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N_p(0, \sigma^2 D_1), \quad D_1 = \text{diag}(\tau_1^2, \ldots, \tau_p^2),
\]

\[
\pi(\sigma^2) = \begin{cases} \frac{1}{\sigma^2} & \text{or} \\ \frac{r^p}{\Gamma(p)} \left( \frac{1}{\sigma^2} \right)^{r-1} \exp\left( -r \frac{1}{\sigma^2} \right), & \sigma, r > 0 \end{cases}
\]

\[
\pi(\tau_1^2, \ldots, \tau_p^2) = \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left( - \frac{\lambda^2 \tau_j^2}{2} \right). \tag{2.6}
\]

For a selection of the tuning parameter \( \lambda \), we may use a cross-validation or information criteria (Konishi and Kitagawa [5]). Park and Casella [8] suggested two approaches, empirical Bayes through marginal maximum likelihood and use of an appropriate hyperprior, in the Bayesian lasso. We consider the following gamma hyperprior on \( \lambda^2 \):

\[
\pi(\lambda^2) = \frac{y^\ell}{\Gamma(\ell)} (\lambda^2)^{\ell-1} \exp\left( y \lambda^2 \right), \quad \lambda^2, y, \ell > 0. \tag{2.7}
\]

### 3 Sparse modifying algorithm

The Bayesian lasso enables us to obtain the posterior distribution. However, there is a problem: the lack of sparsity. Although Park and Casella [8] used the posterior median on point estimation and Hans [4] the posterior mean, both of them are not setting any of the coefficients to zero. This drawback is due to the posterior distribution...
Figure 1: Overview of proposed algorithm
First, substitute zero for some component of estimate vector. And we adopt as the modified estimate with a higher posterior probability in substituted estimate and non-substituted one.

approximated by Gibbs sampler. To overcome this drawback of non-sparsity, we propose an algorithm “sparse modifying algorithm” with the property of setting some coefficients identically equal to zero.

Sparse Modifying Algorithm

Step 1. \( \hat{\beta}_1, \ldots, \hat{\beta}_p \) : Estimated coefficients by Gibbs sampler.
\( \hat{\theta}_1 \) : Estimated parameters vector by hierarchical Bayes estimation. (ex. \( \sigma^2, \tau^2 \))
\( \hat{\theta}_2 \) : Estimated or given parameters vector. (ex. hyper parameter of prior on \( \sigma^2, \lambda \))

Step 2. \( \beta^* = \hat{\beta} \)

Step 3. For \( j = 1, \ldots, p \)
   3-1 \( \beta^*_j = 0 \)
   3-2 if \( g(\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_p, \hat{\theta}_1, \hat{\theta}_2, y, X) \geq g(\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_p, \hat{\theta}_1, \hat{\theta}_2, y, X) \)
   then \( \hat{\beta}_j \leftarrow \beta^*_j \)

where
\[
g(\hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_p, \hat{\theta}_1, \hat{\theta}_2, y, X) := \log f(y|\beta, \theta_1, \theta_2) + \log p(\theta_1, \theta_2) + \log p(\theta_1|\theta_2) \\
(\text{log-likelihood} + \text{log-prior on coefficients} + \text{log-hyperprior})
\]

4 Numerical results
Monte Carlo simulations are conducted to investigate the efficiency of the proposed sparse modifying algorithm focusing on the following points:

a. Efficiency of the proposed algorithm
   How effect for making sparsity and how much influence on prediction accuracy and estimation.

b. Influence of substitution order
   From which coefficient component we should substitute zero in proposed algorithm, and how much influence for different order.

c. Stability
   As shown in Figure 2, the posterior median and posterior mean are more stable than the posterior mode of the Bayesian lasso.

d. Influence of hyperprior
   Park and Casella [8] proposed that \( \sigma^2 \) has the improper prior \( 1/\sigma^2 \). It is, however, hard to choose such prior for the BUGS-applications that enable us the Gibbs sampler, so we use other type of distribution as a prior for \( \sigma^2 \).
Figure 2: Stability of Bayesian Estimates
These are solution path of Bayesian point estimates. Left is the posterior mode, center is the posterior median, and right is the posterior mean.

References


