SPARSE REGRESSION MODELING VIA THE MAP BAYESIAN LASSO

Hoshina, Ibuki
Department of Mathematics, Graduate School of Science and Engineering, Chuo University

https://doi.org/10.5109/1909523

出版情報：Bulletin of informatics and cybernetics. 47, pp.37-58, 2015-12. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 47, 2015

SPARSE REGRESSION MODELING VIA THE MAP BAYESIAN LASSO

By

Ibuki HOSHINA*

Abstract

Sparse regression procedures that are typified by the lasso enable us to perform variable selection and parameter estimation simultaneously. However, the lasso does not give the estimate of error variance, and also the tuning parameter selection still remains an important issue. On the other hand, although the Bayesian lasso can determine the estimate of error variance and the value of a tuning parameter as some Bayesian point estimates, it is difficult to derive sparse solution for the estimates of regression coefficients. To overcome these drawbacks, we propose a MAP Bayesian lasso by using the Monte Carlo integration for the posterior approximation. Monte Carlo simulations and real data examples are conducted to examine the efficiency of the proposed procedure.

Key Words and Phrases: Lasso, tuning parameter estimation, posterior distribution, Monte Carlo integration, Newton's method.

1. Introduction

Computer and sensor technology advancements enable us to get and save the high-dimensional or complex data, and the statistical modeling helps us to obtain some knowledge from such data. The linear regression modeling is used to model a relationship between a response variable and several explanatory variables, and it enables us to predict and interpret mechanisms of phenomena. Parameter estimation and variable selection are fundamentally important in the linear regression modeling. The parameters are usually estimated by using the ordinary least squares or maximum likelihood procedures. Variable selection follows the best subset selection based on model selection criteria such as the AIC (Akaike, 1973) and the BIC (Schwarz, 1978). Cross-validation is also widely used as a model selection criterion. For model selection criteria, we refer to Konishi and Kitagawa (2008). For high-dimensional regression, however, these procedures yield models with poor prediction accuracy. Least square procedures often yield model estimates with large variances, especially when there is a problem of multicollinearity. The best subset selection is often unstable because of its inherent discreteness (Breiman, 1996).

For these drawbacks, one of promising techniques is the lasso (least absolute shrinkage and selection operator) proposed by Tibshirani (1996). The lasso tends to shrink some regression coefficients toward exactly zero by imposing an L_1 penalty on regression coefficients, and does both continuous shrinkage and variable selection simultaneously.

* Department of Mathematics, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. tel +81–(0)3–3817–1745 ibu@gug.math.chuo-u.ac.jp

For the last 20 years, various sparse regression procedures inspired by the lasso have been proposed; e.g. SCAD (smoothly clipped absolute deviation; Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive lasso (Zou, 2006), the group lasso (Yuan and Lin, 2006), and the MCP (minimax concave penalty; Zhang, 2010).

In sparse regression modeling, the selection of adjusted tuning parameters including regularization parameters is a crucial issue, since these procedures depend on values of tuning parameters that identify a set of variables included in a model and also control the bias-variance trade-off in resulting estimates. In the lasso, the degrees of freedom (e.g. Ye, 1998; Efron, 1986; Efron, 2004) quantifies the model complexity and plays a key role in such problem. Efron (2004) showed that Mallows' C_p type criteria (Mallows, 1973) is an unbiased estimator of the true prediction error when degrees of freedom is given, and it often provides better accuracy than cross-validation. It is, however, difficult to derive a closed form of the degrees of freedom of the lasso, so the estimation procedures have been integrated by Zou *et al.* (2007), Kato (2009), Tibshirani and Taylor (2012) and Hirose *et al.* (2013). Especially Zou *et al.* (2007) showed that the number of non-zero regression coefficients is an unbiased estimator of the degrees of freedom of the lasso.

Tibshirani (1996) indicated the relationship between the lasso and Bayesian models: the lasso estimates can be interpreted as a MAP (maximum a posteriori) estimates when the regression coefficients have independent and identical Laplace prior and the likelihood is taken to be a normal linear regression model. The Bayesian lasso (Park and Casella 2008, Hans 2009) is a fully Bayesian analysis, and they suggested the Gibbs sampling for the lasso with the Laplace prior in the hierarchical model. The Bayesian lasso provides the Bayesian credible intervals of the lasso, and it guides the variable selection.

Compared to non-Bayesian modeling, the Bayesian lasso also has two advantages:

1. estimating error variance.
2. choosing the values of tuning parameters.

In the lasso, the estimate of error variance is not directly obtained, and efficient procedures were studied (see e.g. Reid *et al.*, 2014). On the other hand, the Bayesian lasso determines it as mode, median, or mean of the posterior distribution. Tuning parameters which can be viewed as the Bayesian hyper parameters, are estimated by a hierarchical or an empirical Bayesian method.

The Bayesian lasso has two drawbacks: it is difficult to calculate a posterior mode of the regression coefficients, and the resulting regression coefficients are not sparse. Although the posterior mode of the Bayesian lasso coefficients is equivalent to the lasso estimates, it is difficult to calculate the posterior mode because posterior density function is not differentiable at zero. The kernel density estimation may be applicable for this problem. It is however difficult to calculate a stable posterior mode in high-dimensional density estimation. Furthermore, Park and Casella (2008) indicate that the Bayesian lasso (point) estimates for regression coefficients don't take zero value exactly.

To overcome these drawbacks, we propose a new methodology that approximates the posterior density function of the Bayesian lasso by Monte Carlo integration; estimating the posterior mode by Newton's method, and modifying the resulting estimates of regression coefficients to be sparse along a posterior probability. Note that our methodology is not a Bayesian analysis, but we use Bayesian model for sparse modeling. In the Bayesian analysis, the posterior distribution summarizes knowledge about the unknown

parameters on the basis of observations. The MAP procedures including our methodology, however, only perform point estimation, and they do not provide any measure of uncertainty on the estimated parameters. (e.g., Murphy, 2012). Thus, our procedure is not a Bayesian analysis.

The remainder of this paper is organized as follows: Section 2 briefly describes the Bayesian model of the lasso in the linear regression models. In Section 3, we introduce a new methodology that approximates a posterior density function by Monte Carlo integration, estimates a posterior mode by Newton's method which considers a posterior probability, and derives sparse estimates of regression coefficients. Section 4 presents numerical studies for both artificial and real data sets. Some concluding remarks are given in Section 5.

2. Bayesian lasso

2.1. Lasso

We consider the linear regression model

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an n-dimensional response vector, β_0 is an intercept, $\mathbf{1}_n$ is an n-dimensional vector whose components are all one, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ design matrix, with p-dimensional observations for predictor variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-dimensional error vector which elements have independent and identically distributed according to a normal distribution with mean zero and unknown variance σ^2. Without loss of generality, we assume that the response and predictors are standardized:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad j = 1, \ldots, p.$$

Since the error vector $\boldsymbol{\varepsilon}$ has an n-dimensional normal distribution $N_n(\mathbf{0}, \sigma^2 I_n)$, we have a probability density function for the response vector $\boldsymbol{y}$ in the form

$$p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 \right\},$$

a probability density function of an n-dimensional normal distribution

$$N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n),$$

with mean vector $X\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 I_p$.

This leads to the log-likelihood function

$$\log p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2.$$

Thus, the maximum likelihood estimator (MLE) for the regression coefficients vector $\boldsymbol{\beta}$ on model (1) is defined by

$$\hat{\boldsymbol{\beta}}^{\mathrm{MLE}} = \underset{\boldsymbol{\beta}}{\mathrm{argmax}} \left[-\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \right],$$

and this is equivalent to the ordinary least square estimator (OLS)

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \right].$$

Although OLS and MLE procedures are both usual estimation techniques, these procedures often have large variance when the dimensionality p grows to a large size or some variables have strong correlation between each other. The ordinal methods cannot produce sparse solution for estimates of regression coefficients, and do not enable us to perform regression coefficients estimation and variable selection simultaneously.

The lasso estimates (Tibshirani, 1996) for the regression coefficients are obtained by solving L_1 penalized least square problem

$$\hat{\boldsymbol{\beta}}^{\mathrm{lasso}} := \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2}$$

where λ is a regularization parameter which controls the number of predictor variables. The lasso continuously shrinks the coefficients toward 0 as λ increases. In the case of $\lambda = 0$ and $n > p$, $\hat{\boldsymbol{\beta}}^{\mathrm{lasso}}$ is equivalent to the OLS or MLE, and some coefficients are shrunk to exactly zero when the scale of λ is sufficiently large because of nature of the L_1 penalty. From this property, model selection problem in the lasso is naturally equivalent to the tuning parameter selection problem.

Since the objective function in (2) is not differentiable at $\beta_j = 0$ ($j = 1, \ldots, p$), there are no closed form of the lasso estimates. Hence, in order to obtain estimates of regression coefficients, a number of efficient algorithms are proposed; e.g., the shooting algorithm (Fu, 1998), the LARS (Efron *et al.*, 2004), the GPS algorithm (Friedman, 2008) and the coordinate descent algorithm (Friedman *et al.*, 2010).

The lasso has been much studied from various viewpoints. For example, Zou and Hastie (2005), Zou (2006), Yuan and Lin (2006) extended the L_1 penalty into the "$L_1 + L_2$", "weighted", or "group" penalty. Knight and Fu (2000) and Bühlmann and van de Geer (2011) showed asymptotic properties of the lasso-type estimators. Zou *et al.* (2007), Kato (2009), Tibshirani and Taylor (2012), and Hirose *et al.* (2013) investigated the degrees of freedom of the lasso that plays a key role in model selection problem.

2.2. Bayesian lasso

Several articles (e.g., Tibshirani 1996, Park and Casella 2008) described that the lasso estimates can be interpreted as the maximum a posteriori (MAP) estimates when $\beta_1, \ldots, \beta_p$ have an independent and identical Laplace priors $p(\beta) = (\lambda/2) \cdot \exp(-\lambda|\beta|)$. However, the Laplace distribution sometimes makes it difficult to analyze. For example, if we have a normal likelihood and a normal prior, then we can easily obtain the normal posterior distribution through the completing square techniques. By contrast, we can hardly apply such techniques to the Laplace prior. Andrews and Mallows (1974) showed that the Laplace distribution can be expressed as a scale mixture of the normal

distributions with independent exponentially distributed variables:

$$\begin{aligned}\frac{a}{2}\exp\left(-a|z|\right) &= \int_0^\infty \frac{1}{2\pi s}\exp\left(-\frac{1}{2s}z^2\right)\cdot\frac{a^2}{2}\exp\left(-\frac{a^2}{2}s\right)ds \quad (a>0)\\ &= \int_0^\infty N(z|0,s)\cdot \mathrm{Exp}\left(s|\frac{a^2}{2}\right)ds,\end{aligned}$$

where $\mathrm{Exp}(x|a)$ is a probability density function of an exponential distribution with variable x and rate parameter a. In the Bayesian approach, the scale mixture formulation has a predilection for the hierarchical representation of the full model. Park and Casella (2008) proposed the Gibbs sampling for the lasso from hierarchical representation of the full model:

$$\begin{aligned} p(\boldsymbol{y}|X,\boldsymbol{\beta},\sigma^2) &= N_n(\boldsymbol{y}|X\boldsymbol{\beta},\sigma^2 I_n),\\ p(\boldsymbol{\beta}|\sigma^2,\tau_1^2,\ldots,\tau_p^2) &= N_p(\boldsymbol{\beta}|\mathbf{0}_p,\sigma^2 D),\\ p(\sigma^2) &= \frac{1}{\sigma^2}\ \text{or}\ \mathrm{IG}(\sigma^2|\nu_0,\eta_0),\\ p(\tau_1^2,\ldots,\tau_p^2|\lambda) &= \prod_{j=1}^p \mathrm{Exp}\left(\tau_j^2|\frac{\lambda^2}{2}\right),\end{aligned} \tag{3}$$

where $\mathbf{0}_q$ is a q-dimensional vector whose elements are all 0, $D=\mathrm{diag}(\tau_1^2,\ldots,\tau_p^2)$, $\mathrm{IG}(x|\nu,\eta)$ is the probability density function of an inverse-gamma distribution with variable x, shape parameter ν and rate parameter η. The Bayesian lasso enables us to obtain the Bayesian credible intervals, and we can perform the model selection through these interval estimates. The conditional prior of $\boldsymbol{\beta}$ given σ^2 in the model (3) guarantees a unimodal full posterior, which avoids the slow convergence of the Gibbs sampler. For more details, we refer to Andrews and Marrows (1974) and Park and Casella (2008).

The full model in (3) leads to the following full conditional distributions of $\boldsymbol{\beta}$, σ^2, and $1/\tau_1^2,\ldots,1/\tau_p^2$ when $p(\sigma^2)=1/\sigma^2$:

$$\begin{aligned} p_{\mathrm{full}}(\boldsymbol{\beta}|\boldsymbol{y},X,\sigma^2,\tau_1^2,\ldots,\tau_p^2) &= N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y},\sigma^2A^{-1}),\\ p_{\mathrm{full}}(\sigma^2|\boldsymbol{y},X,\boldsymbol{\beta},\tau_1^2,\ldots,\tau_p^2) &= \mathrm{IG}(\sigma^2|\nu_1,\eta_1)\\ p_{\mathrm{full}}(1/\tau_1^2,\ldots,1/\tau_p^2|\boldsymbol{y},X,\boldsymbol{\beta},\sigma^2,\lambda) &= \prod_{j=1}^p \mathrm{IGauss}(1/\tau_j^2|\mu_j',\lambda'),\end{aligned} \tag{4}$$

where

$$\begin{aligned} A &= X^TX+D^{-1},\\ \nu_1 &= \frac{n+p}{2},\quad \eta_1=\frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})+\boldsymbol{\beta}^TD^{-1}\boldsymbol{\beta}}{2},\\ \mu_j' &= \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}},\quad \lambda'=\lambda^2,\end{aligned}$$

and $\mathrm{IGauss}(x|\nu_1,\eta_1)$ is a probability density function of an inverse-gaussian distribution with variable x ($x>0$), mean ν_1, and shape parameter η_1. If $p(\sigma^2)$ is an inverse gamma prior, then $\nu_1=(n+p+\nu_0)/2$ and $\eta_1=\{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})+\boldsymbol{\beta}^TD^{-1}\boldsymbol{\beta}+\eta_0\}/2$.

Park and Casella (2008) also suggested how to choose the Bayesian lasso tuning parameter λ in Bayesian analysis; considering an empirical Bayes through marginal maximum likelihood and a hierarchical Bayes through gamma priors on λ^2. By generating Gibbs samples according to full conditional distributions (4), we can obtain the estimates of $\boldsymbol{\beta}$ and σ^2 as posterior modes (MAP estimates), medians and means.

We can easily calculate the MAP estimates of σ^2 and λ. It is however difficult to obtain the MAP estimates of $\boldsymbol{\beta}$, since the multivariate kernel density estimation requires high computational cost. The univariate kernel density estimation on this problem becomes in-stable when estimates becomes large, since mean and variance-covariance matrix of full conditional of $\boldsymbol{\beta}$ depend on same parameter A^{-1}. Figure 1 illustrates the regularization paths for the low-dimensional diabetes dataset (Efron *et al.*, 2004) solved by the posterior mode (i.e. MAP estimates) and posterior mean. This figure shows that the MAP estimates is more instable than the posterior mean.

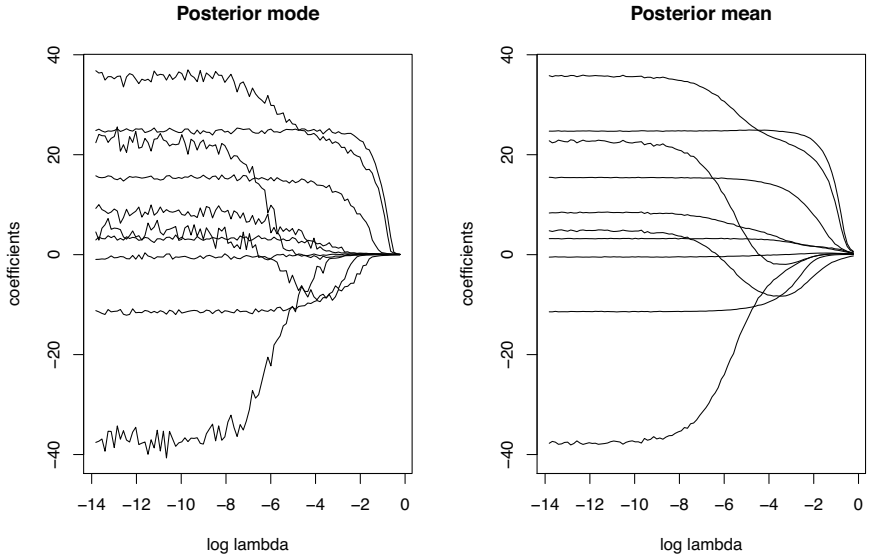

Figure 1: The regularization paths for the low-dimensional diabetes datasets (Efron *et al.*, 2004). Left panel shows the regularization path of the posterior mode, and right panel that for the posterior mean. These point estimates were computed from the Gibbs sampler with size 10,000 after 1,000 burn in iteration.

Moreover, the Bayesian lasso point estimation does not produce some coefficients as exactly zero though the original lasso does. The cause of this problem is from the Gibbs sampler, since there are the estimation errors between the true and estimated posterior mode, and the posterior median and mean do not consist the posterior mode in general.

For this problem, Hoshina (2012) proposed an algorithm that sets some regression coefficients exactly zero so that a posterior probability becomes large. This procedure, however, only corrects for the resulting point estimates to be sparse, and numerically

computed MAP estimates are often instable (Figure 1). Instability of point estimation sometimes leads a poor prediction and estimation accuracy. Hence, we propose another procedure, the MAP Bayesian lasso, in Section 3.

3. Sparse model building methodologies

To obtain the sparse MAP estimates of $\boldsymbol{\beta}$, the optimization methods such as any gradient procedures are required. However, it is difficult to obtain the posterior density function of the Bayesian lasso, and it may be not differentiable at $\beta = 0$ since it includes the Laplace prior. To overcome these drawbacks, we approximate the posterior density by the Monte Carlo integration, and propose a procedure that enables us to obtain the MAP estimates of the Bayesian lasso by Newton's method.

3.1. Posterior distribution approximated by Monte Carlo integration

Since the Bayesian lasso gives us the estimates of σ^2 and λ, our procedure leverages these estimates. Let $\hat{\sigma}^2$ and $\hat{\lambda}$ be the MAP estimates of σ^2 and λ. Then the (conditional) posterior density of $\boldsymbol{\beta}$ given $\hat{\sigma}^2$ and $\hat{\lambda}$ is proportionate to

$$\begin{aligned}
&\int \cdots \int N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \hat{\sigma}^2 I_n) \cdot N_p(\boldsymbol{\beta}|\mathbf{0}_p, \hat{\sigma}^2 D) \left\{ \prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2 \,\middle|\, \frac{\hat{\lambda}^2}{2} \right) \right\} d\tau_1^2 \cdots \tau_p^2 \\
&\quad \propto \int \cdots \int N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \hat{\sigma}^2 A^{-1}) \cdot |D|^{-1/2} \cdot |A|^{-1/2} \\
&\qquad \cdot \exp\left\{ -\frac{1}{2\hat{\sigma}^2} \boldsymbol{y}^T (I_n - XA^{-1}X^T)\boldsymbol{y} \right\} \left\{ \prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2 \,\middle|\, \frac{\hat{\lambda}^2}{2} \right) \right\} d\tau_1^2 \cdots \tau_p^2.
\end{aligned} \tag{5}$$

It is difficult to evaluate the integration in (5) because of complexity of integrand. In general, some approximation methods, such as the Laplace approximation (Tierny and Kadane, 1986), may be used to approximate it. We cannot, however, employ this procedure since the integrand in (5) is not differentiable at $\beta_j = 0$.

In contrast, the Monte Carlo integration is applicable for posterior approximation. The Monte Carlo integration is a well-known numerical technique to approximate a integration in statistics. For example, we often use $\bar{x} = \sum_{m=1}^{M} x_m/M$ as an estimate of the expectation of some random variable X having a probability density function $f(x)$, which can be obtained as

$$\int x f(x) dx \approx \frac{1}{M} \sum_{m=1}^{M} x_m,$$

where $x_1, \ldots, x_M$ is a random sample from the distribution of X. We apply this elementary statistical technique to approximate the integration in (5).

Let $\{\tau^2_{1(m)}, \ldots, \tau^2_{p(m)} : \ m = 1, \ldots, M\}$ be a random sample generated from $\prod_{j=1}^{p}$ $\mathrm{Exp}(\tau_j^2|\hat{\lambda}^2/2)$ artificially, where size M is encouraged to determine sufficiently large

number. Then, we have the following approximation of (5):

$$\frac{1}{M}\sum_{m=1}^{M} N_p(\boldsymbol{\beta}|A_{(m)}^{-1}X^T\boldsymbol{y},\hat{\sigma}^2A_{(m)}^{-1}) \\ \cdot |D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2}\cdot\exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n-XA_{(m)}^{-1}X^T)\boldsymbol{y}\right\}, \tag{6}$$

where $D_{(m)}=\mathrm{diag}(\tau_{1(m)}^2,\ldots,\tau_{p(m)}^2)$, $A_{(m)}=X^TX+D_{(m)}^{-1}$. Since (6) is formed as the sum of differentiable function, (6) is totally differentiable. Hence, the posterior mode of the Bayesian lasso regression coefficients are given by maximizing (6) using Newton's method.

Thus, the approximated posterior distribution $\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y},X,\lambda,\sigma^2)$ and the approximated marginal likelihood $\tilde{p}(\boldsymbol{y}|X,\sigma^2,\lambda)$ of the lasso are respectively given by

$$\begin{aligned}\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y},X,\hat{\sigma}^2,\hat{\lambda}) &= \frac{\frac{1}{M}\sum_{m=1}^{M}N_p(\boldsymbol{\beta}|A_{(m)}^{-1}X^T\boldsymbol{y},\hat{\sigma}^2A_{(m)}^{-1})\cdot\xi_{(m)}}{\int\frac{1}{M}\sum_{\ell=1}^{M}N_p(\boldsymbol{\beta}|A_{(\ell)}^{-1}X^T\boldsymbol{y},\hat{\sigma}^2A_{(\ell)}^{-1})\cdot\xi_{(\ell)}d\boldsymbol{\beta}} \\ &= \sum_{m=1}^{M}\gamma_{(m)}N_p(\boldsymbol{\beta}|A_{(m)}^{-1}X^T\boldsymbol{y},\hat{\sigma}^2A_{(m)}^{-1}), \\ \tilde{p}(\boldsymbol{y}|X,\hat{\sigma}^2,\hat{\lambda}) &= \int\frac{1}{M}\sum_{m=1}^{M}N_p(\boldsymbol{\beta}|A_{(m)}^{-1}X^T\boldsymbol{y},\hat{\sigma}^2A_{(m)}^{-1})\cdot\xi_{(m)}d\boldsymbol{\beta} \\ &= \frac{1}{M}\sum_{m=1}^{M}|D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2} \\ &\qquad\cdot\exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n-XA_{(m)}^{-1}X^T)\boldsymbol{y}\right\},\end{aligned} \tag{7}$$

where

$$\xi_{(m)}=|D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2}\cdot\exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n-XA_{(m)}^{-1}X^T)\boldsymbol{y}\right\},$$
$$\gamma_{(m)}=\frac{\xi_{(m)}}{\sum_{\ell=1}^{M}\xi_{(\ell)}}.$$

Note that, the approximated posterior of the lasso is given in the form of a mixture of normal distributions with mixture weights $\gamma_{(1)},\ldots,\gamma_{(M)}$.

3.2. MAP estimation by Newton's method

Newton's method is one of the second order optimization methods that take the Hessian, i.e. the curvature of the space into account. This iterative algorithm consists of updates of the following form:

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k+\eta_kH_k^{-1}\boldsymbol{g}_k,\quad \boldsymbol{g}_k=\frac{\partial f(\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}},\quad H_k=\frac{\partial^2f(\boldsymbol{\theta}_k)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T},$$

where $\boldsymbol{\theta}_k$ $(k = 1, \ldots)$ is a sequence of variables which converges to the optimal value $\hat{\boldsymbol{\theta}}$, $f(\boldsymbol{\theta})$ is a function which is maximized, and η_k is a step size for k-th update.

In our procedure, the resulting regression coefficients are given by maximizing (6) or $\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, X, \hat{\sigma}^2, \hat{\lambda})$ of (7). We use (6) as the objective function of the maximization problem, and the gradient $\boldsymbol{g}_k$ and the Hessian H_k for k-th update are respectively given as follows:

$$\begin{aligned}
\boldsymbol{g}_k &= \frac{1}{M}(2\pi)^{-p/2}(\hat{\sigma}^2)^{-(p+2)/2} \sum_{m=1}^{M} |D_{(m)}|^{-1/2} \\
&\qquad \cdot \exp\left\{-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta}_k + \boldsymbol{\beta}_k D_{(m)}^{-1}\boldsymbol{\beta}_k)\right\}(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k), \\
H_k &= \frac{1}{M}(2\pi)^{-p/2}(\hat{\sigma}^2)^{-(p+2)/2} \\
&\qquad \cdot \sum_{m=1}^{M} |D_{(m)}|^{-1/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta}_k + \boldsymbol{\beta}_k D_{(m)}^{-1}\boldsymbol{\beta}_k)\right\} \\
&\qquad \cdot \left\{A_{(m)} + \frac{1}{\hat{\sigma}^2}(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k)(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k)^T\right\}.
\end{aligned} \tag{8}$$

We choose the value of step size η_k from candidate values $\{\eta_k^{(1)}, \ldots, \eta_k^{(\ell)}\}$ so that $\boldsymbol{\theta}_{k+1} = \boldsymbol{\beta}_{k+1}$ has the largest posterior density, and we substitute the following function for the posterior density of estimated $\boldsymbol{\beta}$:

$$q(\boldsymbol{\beta}, \boldsymbol{y}, X, \sigma^2, \lambda) = \log N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n) + \sum_{j=1}^{p} \log\left\{\frac{\lambda}{\sqrt{2\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right)\right\}. \tag{9}$$

We use this formula to obtain the MAP estimates of the Bayesian lasso. However, it is difficult to derive sparse solutions for regression coefficients since we use a numerical procedure. For this problem, we can apply the sparse algorithm (Hoshina, 2012), which sets some regression coefficients exactly zero so that a posterior probability becomes large.

Although this procedure enables us to obtain the sparse MAP estimates of the Bayesian lasso, the optimized solution of Newton's method depends on the initial value. Especially, since objective function of this optimization may be waggly, it is considered that many local optimums exist (Figure 2). To avoid this problem, the initial value selection is very important. We employ the posterior means as the initial value of the Newton's method because of its estimation stability, as shown in Figure 1.

The size of numerical integration M may affect the result of our procedure. For this point, an empirical evidence shows that the size of M also suffices at the relatively-small value. Figure 3 shows the solution paths in cases of $M = 50$, 500, 5000 respectively, and all solution paths are similar. From these results, we set M to 500 in numerical studies of Section 4.

We call this procedure the "MAP Bayesian lasso" (Maximum a Approximated Posteriori with the Bayesian lasso). For the details of the procedure, see Algorithm 1.

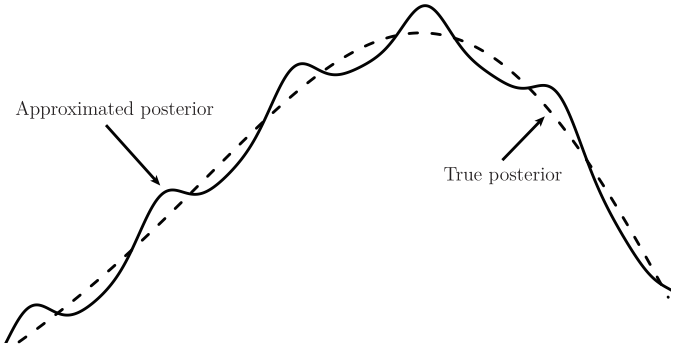

Figure 2: Overview of the objective function of our procedure. Solid and dashed lines illustrate the approximated posterior and true posterior, respectively. Even if true posterior has no local maximum, the approximated posterior may have many local maximums. Thus, it is desired that the initial value of Newton's method is slightly near the global maximum.

Algorithm 1 MAP Bayesian lasso

$\sigma^2 \Leftarrow \hat{\sigma}^2$: posterior mode of σ^2;
$\lambda \Leftarrow \hat{\lambda}$: posterior mode of λ;
Initialize $\boldsymbol{\beta}_0 = \bar{\boldsymbol{\beta}}$: posterior mean;
for $k = 1, 2, \ldots$ until convergence **do**
    Evaluate the gradient $\boldsymbol{g}_k$ of (8);
    Evaluate the Hessian H_k of (8);
    Solve $\boldsymbol{z}_k = H_k^{-1}\boldsymbol{g}_k$;
    for $\ell = 1, 2, \ldots, L$, solve $\boldsymbol{\beta}_{k+1(\ell)} = \boldsymbol{\beta}_k + \eta_{k(\ell)}\boldsymbol{z}_k$ **do**
        Evaluate the value $q_{(\ell)} = q(\boldsymbol{\beta}_{k+1(\ell)}, \boldsymbol{y}, X, \sigma^2, \lambda)$ of (9);
    end for
    $\boldsymbol{\beta}_{k+1} \Leftarrow \underset{\boldsymbol{\beta}_{k+1(\ell)}}{\operatorname{argmax}}\{q_{(\ell)}\}$;
    $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T \Leftarrow \boldsymbol{\beta}_{k+1}$;
end for
$\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p) \Leftarrow \hat{\boldsymbol{\beta}}$;
for $j = 1, 2, \ldots, p$ **do**
    $\tilde{\beta}_j \Leftarrow 0$;
    if $q(\tilde{\boldsymbol{\beta}}, \boldsymbol{y}, X, \sigma^2, \lambda) > q(\hat{\boldsymbol{\beta}}, \boldsymbol{y}, X, \sigma^2, \lambda)$ **then**
        $\hat{\boldsymbol{\beta}} \Leftarrow \tilde{\boldsymbol{\beta}}$;
    else $\tilde{\boldsymbol{\beta}} \Leftarrow \hat{\boldsymbol{\beta}}$;
    end if
end for

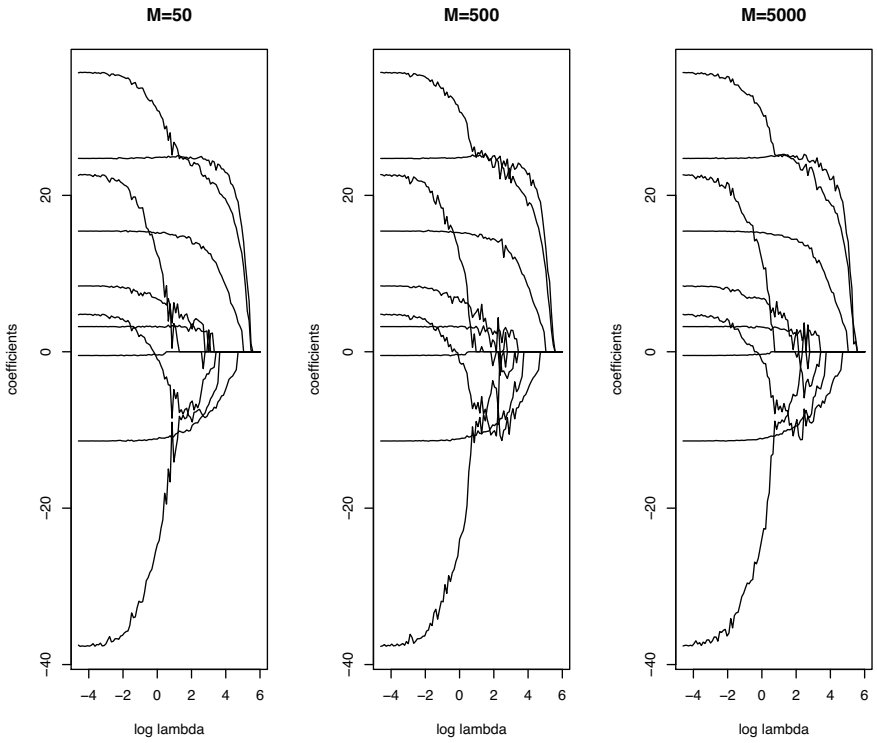

Figure 3: Regularization paths for the diabetes data (Efron *etal.*, 2004) for $M = 50$ (left), $M = 500$ (center) and $M = 5000$ (right).

3.3. Other procedures

This section describes other sparse model building techniques which choose the value of a tuning parameter by model selection criteria.

3.3.1. Baysian lasso with model selection criteria

Suppose that $p(\boldsymbol{y}|\boldsymbol{\theta})$ is a likelihood of n-observation $\boldsymbol{y}$ on parameter $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is a posterior density of $\boldsymbol{\theta}$. Deviance information criterion (DIC) proposed by Spiegelhalter *et al.* (2002) measures the effective number of parameters in a Bayesian model using an information theoretic argument. The measure p_D for parameter $\boldsymbol{\theta}$ is defined by

$$p_D = -2E_{\boldsymbol{\theta}|\boldsymbol{y}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}),$$

where $E_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot)$ denotes the expectation over posterior distribution of $\boldsymbol{\theta}$, and $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$.

Based on this measure, Spiegelhalter *et al.* (2002) proposed a deviance information criterion

$$\mathrm{DIC} = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2p_D.$$

Widely applicable or Watanabe-Akaike information criterion (WAIC) is proposed by Watanabe (2010a, 2010b). WAIC intends to evaluate the model accuracy by the Bayes or Gibbs generalization loss for singular or non-singular model. However, it is difficult to obtain these losses since we need to evaluate a expectation on predictive distribution. For this problem, Watanabe (2010a, 2010b) showed that the consistent estimator of the Bayes generalization loss is given by

$$\begin{aligned}\mathrm{WAIC} = &-\frac{1}{n}\sum_{i=1}^{n}\log E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[p(\boldsymbol{y}|\boldsymbol{\theta})\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left\{E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\left(\log p(y_i|\boldsymbol{\theta})\right)^2\right] - E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\log p(y_i|\boldsymbol{\theta})\right]^2\right\}.\end{aligned}$$

DIC and WAIC need to evaluate the posterior and predictive distribution respectively. The Gibbs sampler enables us to derive these values, and the Bayesian lasso which gives us the Gibbs sample of the lasso can be applicable for these procedures.

3.3.2. Lasso with model selection criteria

The degrees of freedom can lead to several model selection criteria (e.g. Hirose *et al.*, 2013) which may improve prediction accuracy in the lasso.

In the lasso, Zou *et al.* (2007) introduced the AIC (Akaike, 1973), the BIC (Schwarz, 1978) and the Mallows' C_p (Mallows, 1973), respectively, given by

$$\begin{aligned}\mathrm{AIC} &= n\log(2\pi\hat{\sigma}^2) + \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{2\hat{\sigma}^2} + 2\mathrm{DF}, \\ \mathrm{BIC} &= n\log(2\pi\hat{\sigma}^2) + \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{2\hat{\sigma}^2} + \log n\cdot\mathrm{DF}, \\ C_p &= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 + 2\hat{\sigma}^2\mathrm{DF},\end{aligned}$$

where the likelihood of $\boldsymbol{y}$ is given by $N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n)$ and DF is the degrees of freedom of the lasso. Although true value of DF is unknown, Zou *et al.* (2007) showed that the number of non-zero coefficients of the lasso estimate is an unbiased estimator of DF. The AIC and C_p yield the same results when same estimated σ^2 is used.

Hirose *et al.* (2013) also introduced the generalized cross validation (GCV; Craven and Wahba, 1979)

$$\mathrm{GCV} = n\frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{(n - \mathrm{DF})^2}.$$

Note that the GCV does not need estimate of σ^2.

4. Numerical result

In order to examine the effectiveness of our proposed procedure, we conducted Monte Carlo simulations and real data analysis.

4.1. Monte Carlo experiments

Monte Carlo experiments were conducted to investigate the efficacy of our procedure. The data were generated from

$$y = \boldsymbol{x}^T\boldsymbol{\beta}^* + \varepsilon,$$

where $\boldsymbol{\beta}^*$ is a p-dimensional regression coefficients vector, $\varepsilon \sim N(0, \sigma^2)$, and $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is assumed to be a p-variate normal distribution with mean vector $\mathbf{0}_p$. We consider the following cases.

Example 1 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\sigma^2 = 3^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 2 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = 0.85 \cdot \mathbf{1}_p$, $\sigma^2 = 3^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 3 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = (5, \mathbf{0}_{p-1}^T)^T$, $\sigma^2 = 2^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 4 $n = 200$, $p = 40$, $\boldsymbol{\beta}^* = (\mathbf{0}_{10}^T, \mathbf{2}_{10}^T, \mathbf{0}_{10}^T, \mathbf{2}_{10}^T)^T$, $\sigma^2 = 15^2$. $\mathrm{cor}(x_i, x_j) = \rho$ $(i \neq j)$, $\rho = 0.5$.

We compute the following four indicators; prediction squared error (PSE), mean squared error of the regression coefficients vector (MSE), false positive rate (FPR), and false negative rate (FNR) to evaluate the prediction and estimation accuracy of outcome

model, and the simulation results were obtained by 200 Monte Carlo trials.

$$\begin{aligned}
\text{PSE} &= \frac{1}{200}\left(\sum_{k=1}^{200}\|\hat{\boldsymbol{y}}^{(k)}-\tilde{\boldsymbol{y}}^{(k)}\|^2/n\right),\\
\text{MSE} &= \frac{1}{200}\left\{\sum_{k=1}^{200}(\hat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^*)^T R(\hat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^*)\right\},\\
\text{FPR} &= \frac{1}{200}\left(\sum_{k=1}^{200}\#\{\hat{\beta}_j^{(k)}\neq 0\ ;\ \beta_j^*=0\}/\#\{\beta_j^*=0\}\right),\\
\text{FNR} &= \frac{1}{200}\left(\sum_{k=1}^{200}\#\{\hat{\beta}_j^{(k)}= 0\ ;\ \beta_j^*\neq 0\}/\#\{\beta_j^*\neq 0\}\right).
\end{aligned}$$

Where $\hat{\boldsymbol{y}}^{(k)}$ is a predicted vector of k-th data sets, $\tilde{\boldsymbol{y}}^{(k)}$ is a new response vector that independent from $\boldsymbol{y}$, $p\times p$ matrix R is a correlation matrix of $\boldsymbol{x}$, and $\hat{\boldsymbol{\beta}}^{(k)}=(\hat{\beta}_1^{(k)},\ldots,\hat{\beta}_p)^T$ is an estimated regression coefficients vector from k-th data set. We set M in (8) to 500, shape and rate parameter ν_0,η_0 of inverse-gamma prior on σ^2 to both 0.001, and the tuning parameter λ is estimated by the hierarchical Bayesian estimation with non-informative gamma prior on λ^2. In all examples, 3000 samples from the Gibbs sampler were used for estimating parameters after 1000 burn in.

We compare the indicators of our procedure with those of the other procedures described in Section 3.3 and the 10-fold Cross validation (CV). The full Bayesian approach (Mean) which estimates all parameters by the posterior mean is also compared with our procedure. Table 1 shows the comparison of these sparse regression modeling procedures. The result of AIC is not presented, since Mallows' C_p criterion and AIC yield the same results when $\hat{\sigma}^2$ is given. The Bayesian estimates derived by three procedures (Mean, DIC, and WAIC) were calculated by the sparse algorithm (Hoshina, 2012), since they have no sparse solution for the estimates of regression coefficients. The error variance σ^2 was estimated by the MLE in the lasso procedures with C_p and BIC.

The simulation results are summarized as follows:

1. For Examples 1, 3, and 4, the Bayesian procedures except for DIC have smaller errors than all lasso procedures in terms of PSE and MSE.

2. Our procedure has slightly large FPR in Examples 1, 3, 4, but all examples show that our procedure has smaller FNR. This may denotes that our procedure takes in more variables into the estimated model.

3. In Examples 1, 2, and 3, our procedure has the smallest value in terms of PSE, and has the smallest value in terms of MSE in Examples 1, 3.

From the summary of the Monte Carlo simulations, our procedure has better prediction and estimation accuracy. Moreover, it hardly wastes the important variables from the model. Thus, we believe that our proposed methodology seems to be useful in terms of variable selection, parameter estimation and prediction. Note that DIC and WAIC need the Gibbs sampling for each candidate value of λ.

Table 1: Comparison of sparse regression modeling procedures. The values in parenthesis for PSE and MSE are their standard deviations.

Example 1.

	Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
PSE	6.17	8.04	15.18	6.45	7.49	11.66	9.44	12.00
	(2.71)	(4.66)	(6.28)	(3.23)	(4.60)	(9.00)	(7.24)	(5.79)
MSE	3.83	5.57	10.29	4.39	4.25	7.01	5.47	4.67
	(2.89)	(5.35)	(5.28)	(3.37)	(4.02)	(6.96)	(5.90)	(4.90)
FPR	0.53	0.27	0.04	0.46	0.47	0.28	0.39	0.44
FNR	0.09	0.25	0.50	0.14	0.12	0.24	0.18	0.14

Example 2.

	Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
PSE	6.33	8.70	15.30	7.06	6.99	10.71	9.48	12.54
	(2.90)	(5.04)	(6.05)	(3.80)	(4.34)	(7.57)	(6.97)	(5.18)
MSE	4.22	6.12	10.26	4.86	4.21	6.49	5.76	5.30
	(2.28)	(4.29)	(3.38)	(3.21)	(2.84)	(4.40)	(4.16)	(3.92)
FPR	–	–	–	–	–	–	–	–
FNR	0.34	0.55	0.80	0.45	0.36	0.50	0.45	0.42

Example 3.

	Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
PSE	2.59	2.76	6.56	2.79	3.73	6.76	4.64	5.26
	(1.11)	(1.23)	(2.23)	(1.32)	(4.17)	(8.03)	(5.25)	(4.79)
MSE	1.34	1.36	3.37	1.40	1.53	3.81	2.04	1.57
	(1.07)	(1.16)	(2.00)	(1.24)	(3.64)	(7.05)	(4.76)	(3.74)
FPR	0.62	0.44	0.01	0.44	0.42	0.18	0.31	0.35
FNR	0.00	0.00	0.00	0.00	0.02	0.06	0.03	0.02

Example 4.

	Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
PSE	193.70	193.67	437.80	202.37	238.87	315.58	220.94	322.67
	(21.85)	(22.01)	(49.79)	(24.00)	(36.03)	(144.96)	(33.12)	(97.32)
MSE	25.08	24.66	234.72	24.23	67.19	140.80	50.27	106.31
	(5.76)	(5.83)	(46.73)	(7.11)	(34.35)	(137.76)	(27.49)	(100.60)
FPR	0.49	0.42	0.36	0.46	0.28	0.23	0.31	0.26
FNR	0.14	0.15	0.13	0.09	0.26	0.34	0.23	0.31

4.2. Real data analysis

We explore our procedure by the following two types of the diabetes datasets of Efron *et al.* (2004) which have been obtained from 442 diabetes patients.

low-dimensional dataset
: consists of ten baseline variables (age, sex, body mass index, average blood pressure and six blood serum measurements) and the response variable which is a quantitative measure of disease progression one year after baseline.

high-dimensional dataset
: consists of ten baseline variables of low-dimensional dataset and 54 certain interactions. The response variable which is also a quantitative measure of disease progression one year after baseline.

In order to compare the prediction accuracy, the out-of-sample comparison is also conducted. We divide the datasets into 221 training and 221 test data randomly, and we compute the following prediction error for test data after model building in training data for k-th trial:

$$\mathrm{PE}(k) = \|\hat{\boldsymbol{y}}^{(k)}_{\mathrm{train}} - \boldsymbol{y}^{(k)}_{\mathrm{test}}\|^2/221,$$

where $\hat{\boldsymbol{y}}^{(k)}_{\mathrm{train}}$ is a predicted vector of k-th training data, $\tilde{\boldsymbol{y}}^{(k)}_{\mathrm{test}}$ is a response vector of k-th test data. We also set M in (8) to 500, (ν_0, η_0) to (0.001, 0.001), λ is estimated by the hierarchical Bayesian estimation with non-informative gamma prior on λ^2, and 3000 samples from the Gibbs sampler are used for estimating after 1000 burn in, for each datasets.

We compare 8 procedures, the proposed procedure (Proposed), posterior mean (Mean), DIC, WAIC, 10-fold Cross validation (CV), Mallows' C_p (C_p), BIC, and Generalized Cross-validation (GCV).

Table 2 shows the average prediction errors of 50 trials of the out-of-sample comparisons. Table 3 reports the estimated standardized regression coefficients for this datasets, and Figure 4 and 5 also illustrate the estimated standardized regression coefficients.

The results of the real data analysis are summarized as follows:

1. In low-dimensional diabetes dataset, the resulting models of the Bayesian procedures except for DIC have more variables than all lasso procedures. These procedures also have smaller average prediction error.

2. In high-dimensional diabetes datasets, the resulting models of the Bayesian procedures except to a DIC have also more variables than all lasso procedures. Our procedure, posterior mean, and BIC have smaller average prediction error though WAIC has larger value.

From the summary of the real data analysis, our procedure has better prediction accuracy.

Table 2: The average prediction error of the out-of-sample comparison. The values in parenthesis are the standard deviations.

Low-dimensional diabetes dataset

Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
3025.39	3024.16	3856.53	3034.29	4212.28	4397.09	3430.46	4394.36
(203.31)	(207.78)	(268.22)	(205.80)	(993.33)	(1096.13)	(651.82)	(1099.76)

High-dimensional diabetes dataset

Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
3095.19	3090.59	3933.00	3848.07	3152.15	3259.43	3046.11	3230.48
(197.05)	(190.03)	(302.43)	(1597.29)	(259.40)	(386.14)	(184.40)	(327.78)

Table 3: The estimated standardized regression coefficients for low-dimensional diabetes dataset. $*$s in table show variables estimated their coefficients to be exactly zero.

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
Proposed	$*$	-10.62	24.94	14.94	-13.07	3.15	-5.63	5.39	26.65	3.17
Mean	$*$	-10.18	24.96	14.65	-9.80	$*$	-6.73	4.85	25.35	3.07
DIC	$*$	$*$	18.47	3.68	$*$	$*$	$*$	$*$	14.71	$*$
WAIC	$*$	-9.77	24.88	14.34	-7.42	$*$	-7.73	4.34	24.48	2.94
CV	$*$	$*$	17.44	0.24	$*$	$*$	$*$	$*$	14.58	$*$
C_p	$*$	$*$	14.65	$*$	$*$	$*$	$*$	$*$	11.79	$*$
BIC	$*$	$*$	14.65	$*$	$*$	$*$	$*$	$*$	11.79	$*$
GCV	$*$	$*$	14.65	$*$	$*$	$*$	$*$	$*$	11.79	$*$

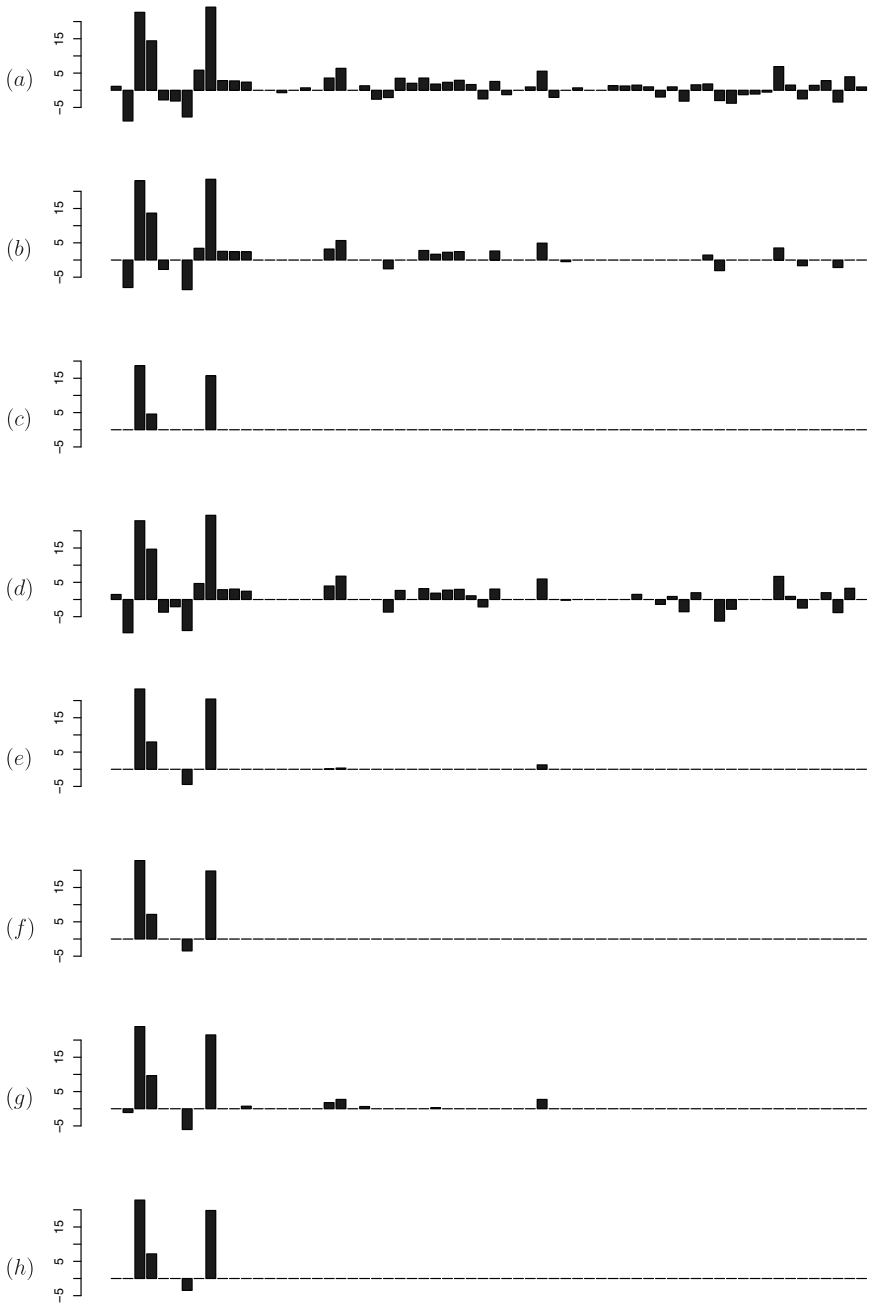

Figure 4: Barplots of the estimated standardized regression coefficients for the high dimensional diabetes dataset: (a) proposed, (b) Mean, (c) the DIC, (d) the WAIC, (e) the CV, (f) the C_p, (g) the BIC, and (h) the GCV.

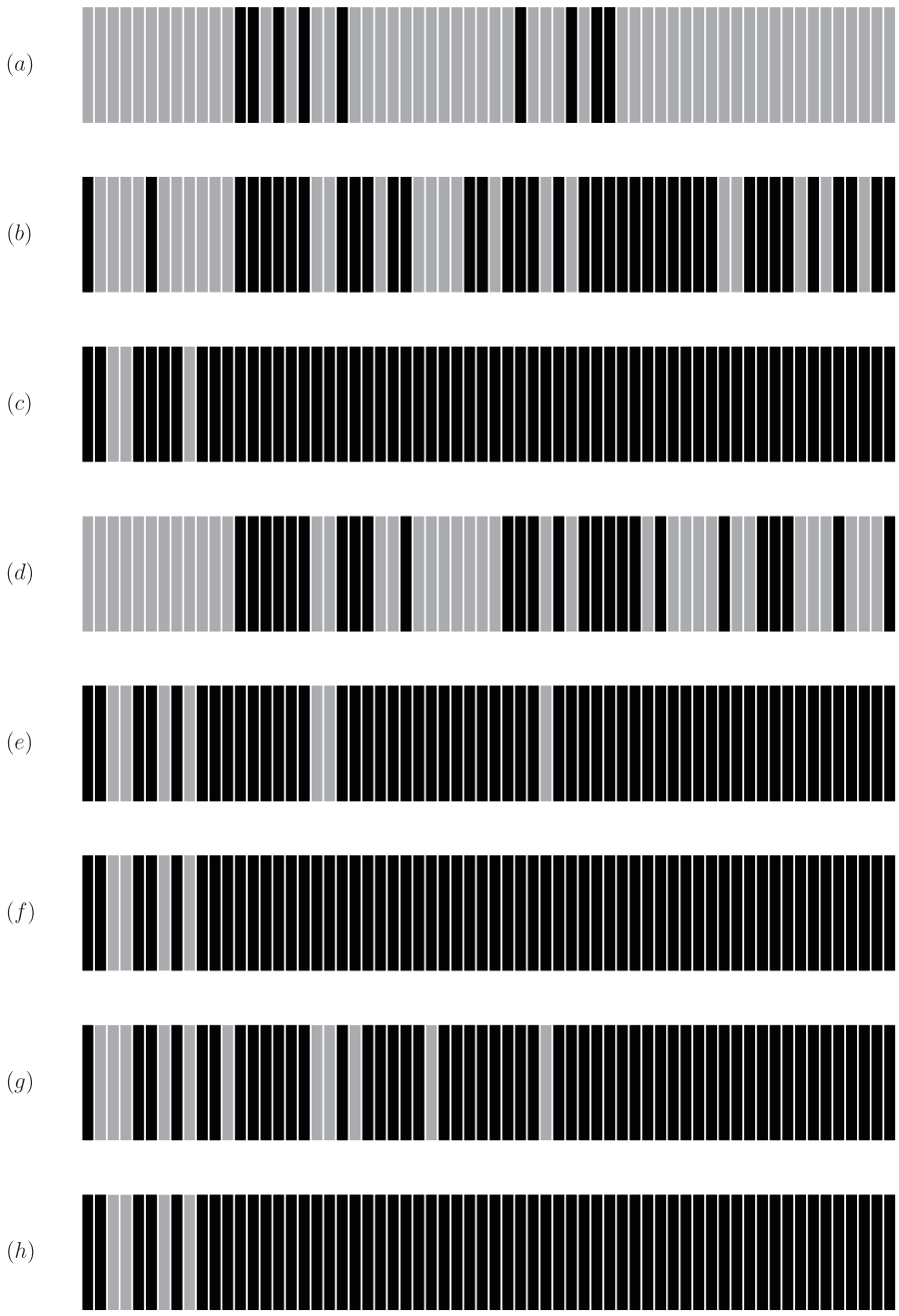

Figure 5: The sparsity of the estimated standardized regression coefficients for the high dimensional diabetes dataset: (a) proposed, (b) Mean, (c) the DIC, (d) the WAIC, (e) the CV, (f) the C_p, (g) the BIC, and (h) the GCV. Grey areas correspond to non-zero coefficients, and black areas correspond to zero coefficients.

4.3. Computational speed

The computational times for parameter estimation and model selection in 8 procedures were compared. Table 4 shows the result of timings for the low-dimensional and high dimensional diabetes datasets in section 4.2. All procedures are computed on a PC with an Intel Core i7 2.8 GHz processor on Mac OSX.

From Table 4, it is shown that WAIC and DIC take a lot of time to perform. Although proposed procedure and posterior mean need more time than 10-fold Cross validation, Mallows' C_p, BIC and Generalized Cross-validation, they are faster than WAIC and DIC in both the low-dimensional and high-dimensional diabetes datasets. Our procedures needs more time to analyze the high-dimensional diabetes datasets than low-dimensional, because we need to calculate an inverse matrix of dimensionality size. Note that the timing of our proposed procedure included the time for computing the posterior mean, since our procedure uses the posterior mean as the initial value.

Table 4: Computational times (seconds) for the low-dimensional and high-dimensional diabetes datasets.

	Proposed	Mean	DIC	WAIC	CV	C_p	BIC	GCV
Low-dim.	2.024	1.588	148.215	340.020	0.080	0.002	0.091	0.003
High-dim.	9.832	5.518	461.248	619.127	0.172	0.005	0.087	0.004

5. Conclusion and remarks

The main aim of the present paper is to investigate model estimation procedures in the Bayesian sparse regression modeling which enables parameter estimation and variable selection simultaneously. We proposed the approximation procedure of the Bayesian lasso posterior by the Monte Carlo integration, and derived the optimization procedure which enables us to obtain sparse MAP estimates of the Bayesian lasso. Monte Carlo experiments showed that our procedure performs well in terms of variable selection, parameter estimation, and prediction. The real data analysis also showed the prediction efficiency of our procedure.

Future studies will be required to consider the generalized sparse regression procedures such as the elastic net, the adaptive lasso, and the group lasso.

Acknowledgements

The author would like to thank the referee for constructive comments and suggestions which significantly improved the manuscript. The author would also like to express my gratitude to Prof. Sadanori Konishi and Prof. Fumitake Sakaori of Chuo University for constant support and encouragement.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principles. *2nd International Symposium on Information Theory*, 267–281.

Andrews, D. L. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, Ser. B, 36, 99–102.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24, 2350–2383.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory, and Applications*, New York: Springer.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377–403.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81, 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99, 619–632.

Efron, B., Hasite, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Friedman, J. (2012). Fast Sparse Regression and Classification. *International Journal of Forecasting*, 28, 722–738.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularizaiton paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7, 397–416.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96, 835–845.

Hirose, K., Tateishi, S. and Konishi, S. (2013). Tuning parameter selection in sparse regression modeling. *Computational Statistics and Data Analysis*, 59, 28–40.

Hoshina, I. (2012). Sparse regression modeling via the Bayesian lasso (in Japanese). *Bulletin of the Computational Statistics of Japan*. 25, 73–85.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, 100, 1338–1352.

Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91, 27–43.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, New York: Springer.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 28, 1356–1378.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15, 661–675.

Murphy, K. (2012). *Machine Learning – a Probabilistic Perspective*, MIT Press.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

Reid, S., Tibshirani, R. and Friedman, J. (2014). A study of error variance estimation in lasso regression. *arXiv*, 1311.5274v2.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40, 1198–1232.

Tierney, L. and Kadane J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

Watanabe, S. (2010a). Equations of states in singular statistical estimation. *Neural Networks*. 23, 20–34.

Watanabe, S. (2010b). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*. 11, 3571–3594.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93, 120–131.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Ser. B, 68, 49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35, 2173–2192.

Received June 30, 2015
Revised November 23, 2015