PREDICTIVE MODEL SELECTION CRITERIA FOR BAYESIAN LASSO REGRESSION

Shuichi Kawano∗, Ibuki Hoshina†, Kaito Shimamura† and Sadanori Konishi‡

ABSTRACT

We consider the Bayesian lasso for regression, which can be interpreted as an $L_1$ norm regularization based on a Bayesian approach when the Laplace or double-exponential prior distribution is placed on the regression coefficients. A crucial issue is an appropriate choice of the values of hyperparameters included in the prior distributions, which essentially control the sparsity in the estimated model. To choose the values of tuning parameters, we introduce a model selection criterion for evaluating a Bayesian predictive distribution for the Bayesian lasso. Numerical results are presented to illustrate the properties of our sparse Bayesian modeling procedure.

1. Introduction

In regression analysis, variable selection plays an important role in the extraction of information from huge-scale datasets with complex structures. Sparse regularization methods, such as lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), the minimax concave penalty (Zhang, 2010), and the seamless $L_0$ penalty (Dicker et al., 2013), have been widely used to perform simultaneous variable selection and parameter estimation. It is well known that penalty functions in regularization correspond to the specification of prior distributions in Bayesian models; that is, the regularization methods can be regarded as a Bayesian approach. For example, coefficient estimates for lasso could be interpreted as a posterior mode under independent Laplace prior distributions. From this perspective, Park and Casella (2008) proposed the Bayesian lasso, which is a Bayesian treatment for lasso. Subsequently, several variations and extensions of the Bayesian lasso have been proposed and used in various fields (e.g., Kabán, 2007; Kyung et al., 2010; Li and Lin, 2010; Mutshinda and Sillanpää, 2010; Griffin and Brown, 2011; Li et al., 2011; Biswas and Lin, 2012; Leng et al., 2014).

A crucial issue for the Bayesian lasso is the choice of values of hyperparameters included in prior distributions, which corresponds to the selection of values of tuning parameters included in the regularization term. Park and Casella (2008) took a hierarchical or empirical Bayes approach to determine the values of hyperparameters. Hans (2010) proposed a variable selection procedure that can treat model uncertainty based on the marginal likelihood. The deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), which is

∗Graduate School of Information Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan E-mail: skawano@ai.is.uec.ac.jp
†Department of Mathematics, Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
‡Department of Mathematics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Key words: Bayesian predictive distribution; Information criterion; Kullback-Leibler information; $L_1$ regularization; Markov chain Monte Carlo
one of the most popular model selection criteria from the viewpoints of Bayesian approaches, and other Bayesian variable selection methods (e.g., O’Hara and Sillanpää, 2009) would also be applicable for addressing the problem. While these methods are mainly obtained by evaluating a marginal likelihood or a plug-in predictive distribution, we take another approach that evaluates a Bayesian predictive distribution. We consider selecting the values of hyperparameters in terms of Bayesian prediction. In Bayesian statistics, it is natural to adopt a Bayesian predictive distribution when one has an interest in prediction. However, there are few studies about choosing the values of hyperparameters by evaluating the Bayesian predictive distribution in the framework of Bayesian lasso regression.

In this paper, we present a model selection criterion by evaluating the Bayesian predictive distribution for the Bayesian lasso from the viewpoints of an information-theoretic approach. First, we obtain an approximated prior distribution by approximating the Laplace prior distribution by the normal prior distribution based on the Kullback-Leibler information. Note that we employ the approximated prior distribution only when a model selection criterion is derived. By using this approximated distribution, we can analytically obtain the formula of the Bayesian predictive distribution for the Bayesian lasso. To derive our proposed criterion, we employ the idea of the predictive information criterion (PIC) proposed by Kitagawa (1997). Several numerical studies are conducted to investigate the performance of our proposed procedure.

The rest of this paper is organized as follows. Section 2 describes the Bayesian lasso. In Section 3, we introduce a model selection criterion for evaluating the Bayesian predictive distribution for the Bayesian lasso. Monte Carlo simulations and real data analysis are conducted to examine the performance of our proposed procedure and to compare it with existing methods in Section 4. Concluding remarks are given in Section 5.

2. Bayesian lasso

2.1. Preliminaries

We consider the linear regression model

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

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is an $n$-dimensional response vector, $\mathbf{1}_n$ is an $n$-dimensional vector whose elements are all one, $X = (x_1, \ldots, x_n)^T$ is an $n \times p$ design matrix, $\beta_0$ is an intercept, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional coefficient vector and $\mathbf{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector distributed as $N_n(\mathbf{0}, \sigma^2 I_n)$. Here, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ ($i = 1, \ldots, n$) denotes a $p$-dimensional covariate vector, $\sigma^2$ is an unknown parameter and $I_n$ is an $n \times n$ identity matrix. Without loss of generality, we assume that the response vector is centered and that the design matrix $X$ is standardized;

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

From this assumption, Equation (1) is replaced by

$$\mathbf{y} = X \beta + \mathbf{\varepsilon}.$$  

Since the error vector $\mathbf{\varepsilon}$ is distributed as a multivariate normal distribution $N_n(\mathbf{0}, \sigma^2 I_n)$, we have a probability density function for the response $y_i$ ($i = 1, \ldots, n$) in the form

$$f(y_i | \mathbf{x}_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{(y_i - \mathbf{x}_i^T \beta)^2}{2\sigma^2} \right], \quad i = 1, \ldots, n.$$
This leads to the log-likelihood function
\[
\sum_{i=1}^{n} \log f(y_i|x_i; \beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.
\] (2)

Hereinafter, we represent the probability density function \( f(y_i|x_i; \beta, \sigma^2) \) as \( f(y_i|\beta, \sigma^2) \), for simplicity.

2.2. Lasso

We estimate the parameters \( \beta \) and \( \sigma^2 \) by maximizing the log-likelihood function with \( L_1 \) penalty:
\[
\max_{\beta, \sigma^2} \left\{ \sum_{i=1}^{n} \log f(y_i|\beta, \sigma^2) - n\lambda \sum_{j=1}^{p} |\beta_j| \right\},
\] (3)

where \( \lambda (>0) \) is a regularization parameter. This estimation procedure is called lasso (Tibshirani, 1996).

Formula (3) is a concave optimization problem, and hence there exists a unique solution for these parameters. Meanwhile, the solution of lasso is not usually expressed in a closed form since the second term in (3) is non-differentiable. Therefore, various estimation algorithms for lasso have been developed to estimate the parameter \( \beta \) numerically, e.g., the LARS (least angle regression) algorithm by Efron et al. (2004) and the coordinate descent algorithm by Friedman et al. (2010). The parameter \( \sigma^2 \) is estimated by \( n^{-1} \sum_{i=1}^{n} (y_i - X\hat{\beta})^T (y_i - X\hat{\beta}) \), where \( \hat{\beta} \) is a lasso estimator of the parameter \( \beta \).

2.3. Bayesian lasso

Although lasso is an efficient method, it cannot provide interval estimates for parameters estimated by exactly zero. To overcome this problem, Park and Casella (2008) considered lasso from a Bayesian viewpoint; specifying prior distributions for the coefficient parameter \( \beta \) and variance \( \sigma^2 \). This procedure is called the Bayesian lasso.

To treat lasso in terms of Bayesian theory, we assume a conditional Laplace prior on \( \beta \) of the form
\[
\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{n\lambda}{2\sqrt{\sigma^2}} \exp \left[ -\frac{n\lambda|\beta_j|}{\sqrt{\sigma^2}} \right]
\] (4)

and the noninformative scale-invariant marginal prior \( \pi(\sigma^2) \propto 1/\sigma^2 \) or inverse-gamma prior \( \pi(\sigma^2) = IG(\nu_0/2, \eta_0/2) \) on \( \sigma^2 \), where \( \nu_0/2 \) is a shape parameter, \( \eta_0/2 \) is a scale parameter, and both of these parameters are positive. Conditioning on \( \sigma^2 \) for the prior distribution is crucial, because it guarantees a unimodal full posterior distribution (see Appendix A in Park and Casella (2008)). As an alternative specification for the prior distribution (4), Park and Casella (2008) proposed the hierarchical representation
\[
\pi(\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2) = \prod_{j=1}^{p} \frac{n}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp \left[ -\frac{n^2\beta_j^2}{2\sigma^2\tau_j^2} \right],
\] (5)

\[
\pi(\tau_1^2, \ldots, \tau_p^2) = \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp \left[ -\frac{\lambda^2\tau_j^2}{2} \right].
\] (6)
The specification of the prior distributions is based on representing the Laplace distribution as a scale mixture of normals:

\[
\pi(\beta|\sigma^2) = \int_0^\infty \cdots \int_0^\infty \pi(\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2) \pi(\tau_1^2, \ldots, \tau_p^2) d\tau_1^2 \cdots d\tau_p^2.
\]

For more details, we refer to Andrews and Mallows (1974) and Park and Casella (2008).

The formulation enables us to implement the Gibbs sampler for \(\beta, \sigma^2\) and \(\tau_1^2, \ldots, \tau_p^2\). Assuming an inverse-gamma prior distribution \(\pi(\sigma^2) = IG(\nu_0/2, \eta_0/2)\) on \(\sigma^2\), the full conditional posterior distributions of \(\beta, \sigma^2\) and \(1/\tau_j^2 \ (j = 1, \ldots, p)\) are, respectively, given by

\[
\begin{align*}
\beta | y, X, \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N_p(A^{-1}X^Ty, \sigma^2A^{-1}), \\
A &= X^TX + n^2D_{\tau}^{-1}, \quad D_{\tau} = \text{diag}(\tau_1^2, \ldots, \tau_p^2), \\
\sigma^2 | y, X, \beta, \tau_1^2, \ldots, \tau_p^2 &\sim IG\left(\frac{\nu_1}{2}, \frac{\eta_1}{2}\right), \\
\nu_1 &= n + p + \nu_0, \quad \eta_1 = \|y - X\beta\|^2 + n^2\beta^T D_{\tau}^{-1}\beta + \eta_0, \\
\frac{1}{\tau_j^2} | \beta_j, \sigma^2, \lambda &\sim IGauss(\mu', \lambda'), \\
\mu' &= \sqrt{\frac{\lambda^2\sigma^2}{n^2\beta_j^2}}, \quad \lambda' = \lambda, \quad j = 1, \ldots, p,
\end{align*}
\]

where \(IGauss(\mu, \lambda)\) denotes the inverse Gaussian distribution with density function

\[
f(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x - \mu)^2}{2\mu^2x}\right], \quad x > 0.
\]

By generating Markov chain Monte Carlo (MCMC) samples according to these posterior distributions, we can obtain the estimates of parameters \(\beta\) and \(\sigma^2\) numerically.

While the Bayesian lasso enables us to treat lasso from a Bayesian viewpoint, a crucial problem still arises, namely, the Bayesian lasso with MCMC lacks sparsity in regression coefficient parameters. Since MCMC is not an optimization algorithm, it cannot estimate the posterior mode. In addition, the posterior means and medians that are estimated by MCMC would not be identical to the posterior mode on finite MCMC samples. To overcome this problem, Hoshina (2012) proposed the sparse algorithm (SA) for the Bayesian lasso. An outline of SA is given as follows. After the MCMC process, the SA gives zero values for some components of the estimated coefficient vector such that the posterior probability becomes large. For the details of the algorithm, see Table 1.

After estimating the parameters by numerical algorithm, we obtain a statistical model \(f(y|\hat{\beta}, \hat{\sigma}^2)\), where \(\hat{\beta}\) and \(\hat{\sigma}^2\) are the estimated parameters. The statistical model depends on values of the hyperparameter \(\lambda\) included in the prior distribution. To select the optimal value of the hyperparameter, we derive a model selection criterion by evaluating the Bayesian predictive distribution for the Bayesian lasso. The criterion is based on the information criterion proposed by Kitagawa (1997).

**Remark** Hyperparameters to be determined include \(\lambda\) in the prior distribution on \(\beta\) and \(\nu_0, \eta_0\) in the prior distribution on \(\sigma^2\). In this paper, we focus on the selection of the value of only hyperparameter \(\lambda\), since it is difficult to optimize values of all hyperparamters \((\lambda, \nu_0, \eta_0)\) simultaneously in terms of computational times. We leave this problem as a future research topic.
Sparse algorithm

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ be a vector of estimates via Bayesian lasso

$\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T \leftarrow \hat{\beta}$

For $j = 1, \ldots, p$

$\tilde{\beta}_j \leftarrow 0$

if $g(\tilde{\beta}, \xi, y) \geq g(\hat{\beta}, \hat{\xi}, y)$ then

$\hat{\beta}_j \leftarrow \tilde{\beta}_j$

end if


3. Model selection criterion

Kitagawa (1997) proposed the predictive information criterion (PIC) for evaluating the Bayesian predictive distribution. A Bayesian predictive distribution is, in general, given by

$$ h(z|y) = \int f(z|\beta, \sigma^2) \pi(\beta, \sigma^2|y) d\beta d\sigma^2, \quad (7) $$

where $z = (z_1, \ldots, z_n)^T$ is an $n$-dimensional future observation, $f(z|\beta, \sigma^2) = \prod_{i=1}^n f(z_i|\beta, \sigma^2)$, and $\pi(\beta, \sigma^2|y)$ is the joint posterior distribution

$$ \pi(\beta, \sigma^2|y) = \frac{f(y|\beta, \sigma^2) \pi(\beta, \sigma^2)}{\int f(y|\beta, \sigma^2) \pi(\beta, \sigma^2) d\beta d\sigma^2}. $$

Using the Bayesian predictive distribution, Kitagawa (1997) derived the PIC in the form

$$ \text{PIC} = -2 \log h(y|y) + 2 B_p, $$

where $B_p$ is the bias term given by

$$ B_p = E_{q(y)} \left[ \log h(y|y) - E_{q(z)} \left[ \log h(z|y) \right] \right] \quad (8) $$

with $q(\cdot)$ being the true distribution that generates the data.

In order to derive PIC for the Bayesian lasso, we obtain the Bayesian predictive distribution (7). In the Bayesian lasso, the prior distribution is formulated by (4). It is, however, difficult to obtain the predictive distribution $h(z|y)$ based on these priors analytically, since we cannot analytically represent the form of the posterior distribution. This problem arises from the fact that the prior distribution $\pi(\beta|\sigma^2)$ is not a conjugate prior for the likelihood function. In Section 3.1, we approximate the prior distribution $\pi(\beta|\sigma^2)$ by a conjugate prior distribution (a normal prior distribution) for the likelihood function.
3.1. Approximated prior distribution

Let \( f(\beta) \) be the Laplace distribution
\[
f(\beta) = \frac{n\lambda}{2\sqrt{\sigma^2}} \exp \left( -\frac{n\lambda|\beta|}{\sqrt{\sigma^2}} \right),
\]
and \( g(\beta|\alpha^2) \) be the normal distribution
\[
g(\beta|\alpha^2) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp \left( -\frac{\beta^2}{2\alpha^2} \right),
\]
where \( \alpha \) is positive.

Our aim is to find the normal distribution that is the closest to the Laplace distribution. Here, we measure the closeness between the distributions in terms of the Kullback-Leibler information (Kullback and Leibler, 1951). We determine the normal distribution \( g(\beta|\hat{\alpha}^2) \), where \( \hat{\alpha}^2 \) is an estimator of \( \alpha^2 \), such that the Kullback-Leibler information between the distributions \( f(\beta) \) and \( g(\beta|\alpha^2) \):
\[
\text{KL}(f, g) = \int_{-\infty}^{\infty} f(\beta) \log \frac{f(\beta)}{g(\beta|\alpha^2)} d\beta
\]
is minimized with respect to the parameter \( \alpha^2 \). The result is given in the following theorem.

**Theorem 1** The minimum of the Kullback-Leibler information attains at \( \hat{\alpha}^2 = \left( \frac{\sqrt{\sigma^2}}{n\lambda} \right)^2 \).

**Proof.** The Kullback-Leibler information between \( f(\beta) \) and \( g(\beta|\alpha^2) \) is calculated as
\[
\text{KL}(f, g) = \log(n\lambda) - \log(2\sqrt{\sigma^2}) + \frac{1}{2} \log(2\pi\alpha^2) - 1 + \frac{1}{\alpha^2} \left( \frac{\sqrt{\sigma^2}}{n\lambda} \right)^2.
\]
A minimizer of (12) is \( \hat{\alpha}^2 = \frac{2(\sqrt{\sigma^2}/n\lambda)^2}{1+\frac{1}{\alpha^2}} \), which is obtained by solving the equation \( \partial\text{KL}(f, g)/\partial\alpha^2 = 0 \).

From this result, the Laplace distribution \( f(\beta) \) can be approximated by the normal distribution \( g(\beta|\hat{\alpha}^2) \), and we have
\[
\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{n\lambda}{2\sqrt{\sigma^2}} \exp \left( -\frac{n\lambda|\beta_j|}{\sqrt{\sigma^2}} \right) \approx \tilde{\pi}(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{n\lambda}{\sqrt{2\pi(2\sigma^2)}} \exp \left( -\frac{n^2\lambda^2 \beta_j^2}{2(2\sigma^2)} \right).
\]
The approximated distribution \( \tilde{\pi}(\beta|\sigma^2) \) can be regarded as the closest to the Laplace distribution \( \pi(\beta|\sigma^2) \) in terms of the Kullback-Leibler information. Note that the approximated distribution is employed only when we obtain a model selection criterion, and that the Laplace distribution is employed when we estimate the coefficient parameters.

**Remark** As a measure of the closeness between the Laplace distribution (9) and the normal distribution (10), we considered the Kullback-Leibler information. We can, however, employ various measures of the closeness instead of the Kullback-Leibler information.

For example, the \( L^2 \) distance
\[
L^2(f, g) = \int_{-\infty}^{\infty} \left\{ f(\beta) - g(\beta|\alpha^2) \right\}^2 d\beta
\]
is available as a measure of the closeness. The quantity is calculated as
\[
L^2(f, g) = \frac{n\lambda}{4\sqrt{\sigma^2}} - \frac{2n\lambda}{\sqrt{\sigma^2}} \exp \left( -\frac{\alpha^2 n^2 \lambda^2}{2\sigma^2} \right) \Phi \left( -\sqrt{\frac{\alpha^2 n^2 \lambda^2}{\sigma^2}} \right) + \frac{1}{2\sqrt{\pi\alpha^2}},
\]
where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Then, the estimator $\hat{\sigma}^2$ is obtained by solving $\partial L^2(f,g)/\partial \sigma^2 = 0$, that is,

$$
\frac{1}{4\sqrt{\pi(\alpha^2)^3/2}} + \frac{n\lambda}{\sqrt{\sigma^2}} \left[ \frac{n^2\lambda^2}{\sigma^2} \Phi\left( -\frac{\alpha^2\lambda}{\sigma^2} n\lambda \right) - \frac{n\lambda}{\sqrt{\alpha^2\sigma^2}} \phi\left( -\frac{\alpha^2\lambda}{\sigma^2} n\lambda \right) \right] \exp\left( \frac{\alpha^2 n^2\lambda^2}{2\sigma^2} \right) = 0
$$

with respect to the parameter $\alpha^2$. Here, $\phi(\cdot)$ is the probability distribution function of the standard normal distribution. Since the estimator $\hat{\sigma}^2$ cannot be obtained analytically, we use any numerical algorithms to estimate the parameter.

It is interesting to compare the performance of the approximated distribution based on the Kullback-Leibler information with that based on the $L^2$ distance. In addition, it is difficult to evaluate the approximate accuracy of our approximated prior distribution in terms of both theoretic and numerical analyses, because other approximate methods for the Laplace distribution have not been proposed until now. We will discuss the problems in another paper.

### 3.2. Bayesian predictive distribution for Bayesian lasso

Assuming the approximated prior distribution $\tilde{\pi}(\beta|\sigma^2)$ on $\beta$ and an inverse gamma distribution $\pi(\sigma^2) = IG(\nu_0/2, \eta_0/2)$ on $\sigma^2$ as in Section 2.3, we derive the joint prior distribution $\pi(\beta, \sigma^2)$ of the form

$$
\pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2) \approx \tilde{\pi}(\beta|\sigma^2)\pi(\sigma^2).
$$

From the prior distribution and Bayes’ rule, the joint posterior distribution can be expressed as

$$
\pi(\beta, \sigma^2|y) = \pi_1(\beta|\sigma^2, y)\pi_2(\sigma^2|y),
$$

where each posterior distribution is given by

$$
\pi_1(\beta|\sigma^2, y) = N_p(\hat{\beta}_n, \sigma^2 A_n), \quad \pi_2(\sigma^2|y) = IG\left( \frac{\nu_n}{2}, \frac{\eta_n}{2} \right).
$$

Here,

$$
A_n = \left( X^TX + \frac{n^2\lambda^2}{2} I_p \right)^{-1}, \quad \hat{\beta}_n = A_n X^T y,
$$

$$
\nu_n = n + \nu_0, \quad \eta_n = \eta_0 + y^T \hat{\beta}_n A_n^{-1} \hat{\beta}_n.
$$

Note that if the prior distribution $\pi(\beta|\sigma^2)$ in (4) is used instead of the approximated prior distribution $\tilde{\pi}(\beta|\sigma^2)$, we cannot obtain the posterior distribution $\pi_1(\beta|\sigma^2, y)$.

Using the posterior distributions, we obtain the Bayesian predictive distribution for the Bayesian lasso given by

$$
h(z|y) = \int f(z|\beta, \sigma^2)\pi(\beta, \sigma^2|y)d\beta d\sigma^2
$$

$$
= \frac{\Gamma\left( \frac{n + \nu_n}{2} \right)}{\Gamma\left( \frac{\nu_n}{2} \right)} |\hat{\Sigma}|^{-1/2} \left[ 1 + \frac{1}{\nu_n} (z - X\hat{\beta}_n)^T \hat{\Sigma}^{-1} (z - X\hat{\beta}_n) \right]^{-(n + \nu_n)/2},
$$

where $\hat{\Sigma} = (\hat{\eta}_n/\nu_n) (X A_n X^T + I_n)$ and $\Gamma(\cdot)$ is the Gamma function. This predictive distribution is an $n$-dimensional $t$-distribution with $\nu_n$ degrees of freedom.
3.3. Proposed criterion

Next, we need to calculate the bias term (8), since the Bayesian predictive distribution is obtained by formula (13) in Section 3.2. It is, however, difficult to calculate the bias term analytically, because the Bayesian predictive distribution \( h(z|y) \) in (13) is an \( n \)-dimensional \( t \)-distribution. Hence, we approximate the distribution \( h(z|y) \) by a normal distribution \( f(z|\tilde{\beta}, \tilde{\sigma}^2) \) in the form

\[
h(z|y) = f(z|\tilde{\beta}, \tilde{\sigma}^2) \left\{ 1 + O_p(n^{-1}) \right\},
\]

where \( \tilde{\beta} \) and \( \tilde{\sigma}^2 \) are, respectively, given by

\[
\tilde{\beta} = \left( X^T X + \frac{n^2\lambda^2}{2} I_p \right)^{-1} X^T y,
\]

\[
\tilde{\sigma}^2 = \frac{(y - X\tilde{\beta})^T (y - X\tilde{\beta}) + \frac{n^2\lambda^2}{2} \tilde{\beta}^T \tilde{\beta} + \eta_0}{n + p + \nu_0 + 2}.
\]

This approximation is based on the Laplace approximation (Tierney and Kanade, 1986). For details of this approximation, we refer to Konishi and Kitagawa (2008).

For the approximated predictive distribution \( f(z|\tilde{\beta}, \tilde{\sigma}^2) \) in the equation (14), we define an approximated predictive information criterion (aPIC) as follows:

\[
aPIC = -2 \log h(y|y) + 2B_p^*,
\]

where the approximated bias term \( B_p^* \) is given by

\[
B_p^* = E_q(y) \left[ \log f(y|\tilde{\beta}, \tilde{\sigma}^2) - E_q(z) [\log f(z|\tilde{\beta}, \tilde{\sigma}^2)] \right]
\approx -\frac{1}{2\tilde{\sigma}^2} E_q(y) [(y - X\tilde{\beta})^T (y - X\tilde{\beta}) - E_q(z) [(z - X\tilde{\beta})^T (z - X\tilde{\beta})]] .
\]

Using the results of Kitagawa (1997) and Kim et al. (2012), we can calculate the approximated bias term as

\[
B_p^* \approx \left( \frac{\sigma^*}{\tilde{\sigma}^2} \right) \text{tr} \left[ X \left( X^T X + \frac{n^2\lambda^2}{2} I_p \right)^{-1} X^T \right],
\]

where \( \sigma^* \) is a specific value such that \( q(z) = f(z|\beta^*, \sigma^*) \).

Then we obtain aPIC in the form

\[
aPIC = -2 \log \Gamma \left( \frac{n + \nu_n}{2} \right) + 2 \log \Gamma \left( \frac{\nu_n}{2} \right) + n \log (\pi \nu_n) + \log |\hat{\Sigma}|
+ (n + \nu_n) \log \left[ 1 + \frac{1}{\nu_n} (y - X\hat{\beta}_n)^T \Sigma^{-1} (y - X\hat{\beta}_n) \right]
+ 2 \left( \frac{\sigma^*}{\tilde{\sigma}^2} \right) \text{tr} \left[ X \left( X^T X + \frac{n^2\lambda^2}{2} I_p \right)^{-1} X^T \right].
\]

Since the value of \( \sigma^* \) is generally unknown, we replace \( \sigma^* \) by the mode of the posterior distribution \( \tilde{\sigma}^2 \), and have
Model Selection Criteria for Bayesian Lasso

\[ aPIC = -2 \log \Gamma \left( \frac{n + \nu_n}{2} \right) + 2 \log \Gamma \left( \frac{\nu_n}{2} \right) + n \log(\pi \nu_n) + \log \hat{\Sigma} \\
+ (n + \nu_n) \log \left[ 1 + \frac{1}{\nu_n} (y - X\hat{\beta}_n)^T \hat{\Sigma}^{-1} (y - X\hat{\beta}_n) \right] \\
+ 2 \text{tr} \left[ X \left( X^T X + \frac{n^2 \lambda^2}{2} I_p \right)^{-1} X^T \right]. \]  

(15)

The value of the hyperparameter \( \lambda \) is selected as the minimizer of aPIC in (15).

### 3.4. Other selection methods

#### 3.4.1. Deviance information criterion

Spiegelhalter et al. (2002) introduced a measure for the effective number of parameters in a model from a Bayesian viewpoint, using an information-theoretic argument. The measure is defined by

\[ p_D = -2E_{\pi(\beta,\sigma^2|y)} \left[ \log f(y|\beta,\sigma^2) \right] + 2 \log f(y|\bar{\beta},\bar{\sigma}^2), \]

where \( \bar{\beta} \) and \( \bar{\sigma}^2 \) are the posterior means defined by \( \bar{\beta} = E_{\pi(\beta|y,\sigma^2)}[\beta] \) and \( \bar{\sigma}^2 = E_{\pi(\sigma^2|y)}[\sigma^2] \), respectively.

Based on this measure, Spiegelhalter et al. (2002) proposed a deviance information criterion (DIC) in the form

\[ \text{DIC} = -2 \log f(y|\bar{\beta},\bar{\sigma}^2) + 2p_D. \]

The optimal value of hyperparameter \( \lambda \) is chosen by selecting the one that minimizes the value of DIC. DIC is widely used in various fields of research including statistical science, ecology, and marketing (e.g., Brady et al., 2004; Pieters and Wedel, 2004; Martin et al., 2005; Bolker et al., 2009).

#### 3.4.2. Full Bayesian approach

An alternative method for choosing the hyperparameter \( \lambda \) is a fully Bayesian approach. Park and Casella (2008) assumed the class of gamma prior distributions on \( \lambda^2 \) given by

\[ \text{Gamma} \left( \lambda^2 | r, \delta \right) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp \left[ -\delta \lambda^2 \right], \]  

(16)

where \( r \) and \( \delta \) are adjusted parameters with positive values. In our numerical examples, we set \( r = \delta = 0.001 \).

The specification in (16) has some attractive properties. For example, the prior distribution on \( \lambda^2 \) in (16) enables us to easily implement the Gibbs sampler. For more details of other properties, we refer to Park and Casella (2008).

#### 3.4.3. Widely applicable information criterion

Based on singular learning theory, Watanabe (2010) proposed widely applicable information criteria (WAIC) for evaluating both regular and singular statistical models. WAIC for Bayesian lasso regression is given by
\[ \text{WAIC} = \frac{-1}{n} \sum_{i=1}^{n} \log \left[ E_{\pi(\beta, \sigma^2 | y)} \left[ f(y_i | \beta, \sigma^2) \right] \right] 
+ \frac{1}{n} \sum_{i=1}^{n} \left[ E_{\pi(\beta, \sigma^2 | y)} \left\{ \log f(y_i | \beta, \sigma^2) \right\}^2 \right] - \left\{ E_{\pi(\beta, \sigma^2 | y)} \left[ \log f(y_i | \beta, \sigma^2) \right] \right\}^2. \]

4. Numerical studies

4.1. Monte Carlo simulations

The performances of our proposed method were investigated through simulation studies. We generated data according to the linear regression model

\[ y = \mathbf{x}^T \beta^* + \varepsilon, \quad (17) \]

where \( \beta^* \) is a \( p \)-dimensional true coefficient vector, \( \varepsilon \sim N(0, \sigma^2) \), and \( \mathbf{x} \) was generated from a multivariate normal distribution with mean vector \( \mathbf{0}_p \) and covariance matrix \( \Sigma \). The structure of the covariance matrix is given below. In this simulation, we considered four cases inspired by Tibshirani (1996) as follows:

- **Case 1:** In this case we simulated 200 data sets with 20, 50, or 100 observations. Here, we set \( \beta^* = (3, 1.5, 0, 0, 2, 0, 0)^T \), \( \sigma = 3 \), and the correlation between \( x_i \) and \( x_j \) was \( 0.5^{|i-j|} \).

- **Case 2:** The second case is the same as Case 1, but with \( \beta^* = 0.85 \cdot \mathbf{1}_8 \).

- **Case 3:** The third case is the same as Case 1, but with \( \beta^* = (0.5, \mathbf{0}_7^T)^T \), and \( \sigma = 2 \).

- **Case 4:** In this case we simulated 200 data sets with 50, 100, or 200 observations. Here, we set \( \beta^* = (0_5^T, 2_5^T, 0_5^T, 0.5_5^T)^T \), and \( \sigma = 5 \).

In all cases, 2,000 samples from the MCMC simulation were used for estimating the parameters, where the first 1,000 samples were discarded as burn-in. In addition, we confirmed the convergence of the Markov chain simulations by using R.hat (Gelman and Rubin, 1992); the values were close to one. The hyperparameter \( \lambda \) was tested for 200 values; \( \lambda_i = \lambda_{\text{min}} \cdot \exp[(\log \lambda_{\text{max}} - \log \lambda_{\text{min}}) \cdot (i/200)] \) \( (i = 1, \ldots, 200) \), where \( \lambda_{\text{max}} \) is such that all coefficient parameters are zero and \( \lambda_{\text{min}} \) is \( 10^{-4} \) when \( n = 20 \) and \( 10^{-4}/n \) when \( n \) is larger than 50.

The performances of our proposed procedure were evaluated in terms of three accuracies; variable selection, estimation, and prediction accuracies. As the variable selection accuracy, we employed the true positive rate (TPR), true negative rate (TNR), and true sign rate (TSR), respectively, defined by

\[ \text{TPR} = \frac{1}{200} \sum_{k=1}^{200} \left\{ \frac{j : \hat{\beta}_j^{(k)} \neq 0 \land \beta_j^* \neq 0}{\{ j : \beta_j^* \neq 0 \}} \right\}, \]
\[ \text{TNR} = \frac{1}{200} \sum_{k=1}^{200} \left\{ \frac{j : \hat{\beta}_j^{(k)} = 0 \land \beta_j^* = 0}{\{ j : \beta_j^* = 0 \}} \right\}, \]
\[ \text{TSR} = \frac{1}{200} \sum_{k=1}^{200} \left\{ \frac{\text{sign}(\hat{\beta}_j^{(k)}) = \text{sign}(\beta_j^*)}{p} \right\}, \]
Model Selection Criteria for Bayesian Lasso

where \( \hat{\beta}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^T \) is the estimated coefficient vector for the \( k \)-th data set, and \( |\{\ast\}| \) is the number of elements included in a set \( \{\ast\} \). The estimation and prediction accuracies are determined by MSE and PSE as follows:

\[
\text{MSE} = \frac{1}{200} \sum_{k=1}^{200} (\hat{\beta}^{(k)} - \beta^*)^T \Sigma (\hat{\beta}^{(k)} - \beta^*),
\]

\[
\text{PSE} = \frac{1}{200} \sum_{k=1}^{200} \left\{ \frac{1}{n} \| \hat{y}^{(k)} - \tilde{y}^{(k)} \|_2^2 \right\},
\]

where \( \hat{y}^{(k)} = x^{(k)T} \hat{\beta}^{(k)} \), \( x^{(k)} \) is the predictor for the \( k \)-th data set, and \( \tilde{y}^{(k)} \) is a future observation generated from the true model \( (17) \).

For each case, we compared nine procedures: aPIC (proposed procedure), aPIC + SA (aPIC with the sparse algorithm proposed by Hoshina (2012)), DIC, DIC + SA, Blasso (fully Bayesian procedure for the Bayesian lasso proposed by Park and Casella (2008)), Blasso + SA, WAIC, WAIC + SA, and Lasso. Except for Lasso, the parameters were estimated by using the posterior means, and the values of the hyperparameters \( \nu_0 \) and \( \eta_0 \) involved in the prior distribution on \( \sigma^2 \) were set to 0.001. The tuning parameter in Lasso was selected by 10-fold cross-validation.

Tables 2 and 3 summarize the simulation results. We observe that aPIC has smaller MSE and PSE than other methods in Case 2 when \( n = 20 \) and Case 3 when \( n = 50, 100 \), while aPIC+SA has a larger TNR than other methods in Case 4. DIC or DIC+SA provides slightly smaller TSR than other methods in many cases. While Blasso or Blasso+SA outperforms other methods in Case 3 when \( n = 20 \) with respect to MSE or PSE, these methods tend to have poor performances in other cases. WAIC and WAIC+SA are better than other methods in terms of MSE and PSE in many cases, but WAIC provides the largest MSE and PSE in Case 3 when \( n = 20 \). Lasso provides the largest MSE in Case 3 when \( n = 50 \), although Lasso is competitive with other methods in many cases.

We also compared run-times of the methods: aPIC, DIC, Blasso, and WAIC. Case 1 when \( n = 20 \) was performed two times, and we averaged the computational times. The computational times of DIC, Blasso, and WAIC were 181.47 times, 0.93 times, and 203.74 times as much as aPIC, respectively. From this result, we observe that the computational time of aPIC is competitive with that of Blasso, while DIC and WAIC require much longer computational times compared to aPIC or Blasso.

4.2. Real data examples

By applying our proposed method to real datasets, we examined the effectiveness of our proposed procedure. We used four benchmark datasets: diabetes, Boston housing, Parkinson’s disease, and communities and crimes datasets. The diabetes dataset is available from the lars package in the software R (R Core Team, 2015). Remaining datasets are obtained from UCI database (http://archive.ics.uci.edu/ml/index.html). The numbers of observations and predictors for the four datasets are summarized in Table 4. Note that we deleted missing values for Parkinson’s disease and communities and crimes datasets.

We randomly and equally divided each dataset into training data and test data. Using the training data, we implemented our proposed procedures (aPIC and aPIC+SA), and then computed PSEs by using the test data. We repeated this procedure 200 times. In addition to our proposed procedures, we implemented DIC, DIC+SA, Blasso, Blasso+SA, and Lasso. WAIC and WAIC+SA were not implemented owing to the computational problem (for
Table 2: The results for Case 1 and Case 2.

| Case 1       | Case 2       |
|--------------|--------------|
| n = 20       | n = 20       |
| TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC     | 1.00 | 0.00 | 0.37 | 4.79 | 11.88 | 1.00 | — | 0.93 | 3.42 | 11.16 |
| aPIC+SA  | 0.81 | 0.62 | 0.69 | 5.64 | 12.56 | 0.51 | — | 0.49 | 5.83 | 13.04 |
| DIC      | 1.00 | 0.00 | 0.36 | 5.31 | 11.94 | 1.00 | — | 0.88 | 4.38 | 11.38 |
| DIC+SA   | 0.90 | 0.43 | 0.60 | 5.61 | 12.17 | 0.70 | — | 0.64 | 5.32 | 12.05 |
| Blasso   | 1.00 | 0.00 | 0.37 | 5.06 | 12.16 | 1.00 | — | 0.91 | 3.80 | 11.46 |
| Blasso+SA| 0.65 | 0.73 | 0.70 | 8.10 | 14.49 | 0.45 | — | 0.44 | 6.64 | 13.74 |
| WAIC     | 1.00 | 0.00 | 0.37 | 4.60 | 11.46 | 1.00 | — | 0.88 | 4.04 | 11.20 |
| WAIC+SA  | 0.90 | 0.57 | 0.72 | 4.33 | 11.61 | 1.00 | — | 0.91 | 3.80 | 11.46 |

| n = 50       | n = 50       |
| TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC     | 1.00 | 0.00 | 0.38 | 1.39 | 9.95 | 1.00 | — | 0.99 | 1.43 | 10.42 |
| aPIC+SA  | 0.99 | 0.58 | 0.73 | 1.38 | 9.96 | 0.81 | — | 0.81 | 2.03 | 10.91 |
| DIC      | 1.00 | 0.00 | 0.38 | 1.56 | 10.04 | 1.00 | — | 0.97 | 1.47 | 10.44 |
| DIC+SA   | 0.99 | 0.42 | 0.63 | 1.57 | 10.06 | 0.90 | — | 0.89 | 1.59 | 10.55 |
| Blasso   | 1.00 | 0.00 | 0.38 | 1.42 | 9.98 | 1.00 | — | 0.98 | 1.36 | 10.34 |
| Blasso+SA| 0.98 | 0.51 | 0.69 | 1.65 | 10.16 | 0.86 | — | 0.86 | 1.72 | 10.63 |
| WAIC     | 1.00 | 0.00 | 0.38 | 1.43 | 9.99 | 1.00 | — | 0.97 | 1.41 | 10.00 |
| WAIC+SA  | 1.00 | 0.42 | 0.64 | 1.42 | 10.00 | 0.88 | — | 0.93 | 1.56 | 10.11 |
| Lasso    | 1.00 | 0.56 | 0.72 | 1.34 | 9.94 | 0.91 | — | 0.90 | 1.71 | 10.60 |

| n = 100     | n = 100     |
| TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC     | 1.00 | 0.00 | 0.38 | 0.63 | 9.71 | 1.00 | — | 1.00 | 0.85 | 9.76 |
| aPIC+SA  | 1.00 | 0.50 | 0.69 | 0.61 | 9.70 | 0.96 | — | 0.96 | 0.94 | 9.84 |
| DIC      | 1.00 | 0.00 | 0.38 | 0.66 | 9.74 | 1.00 | — | 0.99 | 0.79 | 9.69 |
| DIC+SA   | 1.00 | 0.47 | 0.67 | 0.64 | 9.73 | 0.98 | — | 0.98 | 0.80 | 9.70 |
| Blasso   | 1.00 | 0.00 | 0.38 | 0.65 | 9.73 | 1.00 | — | 1.00 | 0.76 | 9.67 |
| Blasso+SA| 1.00 | 0.41 | 0.63 | 0.73 | 9.81 | 0.98 | — | 0.97 | 0.79 | 9.69 |
| WAIC     | 1.00 | 0.00 | 0.38 | 0.66 | 9.46 | 1.00 | — | 0.99 | 0.76 | 9.54 |
| WAIC+SA  | 1.00 | 0.50 | 0.69 | 0.65 | 9.45 | 0.97 | — | 0.98 | 0.79 | 9.56 |
| Lasso    | 1.00 | 0.56 | 0.73 | 0.62 | 9.70 | 0.98 | — | 0.98 | 0.82 | 9.71 |

details, memory shortage on our PC). For all datasets, we generated 4,000 MCMC samples, and then the first 1,000 samples were discarded as burn-in. We observed that the MCMC simulations converged, since the R.hat ratios were close to one.

Figure 1 shows the boxplots of the PSEs. Note that we eliminated one result for the communities and crimes dataset, since the result was clearly an outlier. From the figure, we observe that Blasso and Blasso+SA are often superior to other methods, although the Blasso has large variances in the diabetes dataset. Meanwhile, our proposed procedures, aPIC and aPIC+SA, produce small median values of PSEs similar to Blasso and Blasso+SA except for the Parkinson dataset, and have variances that are small and relatively stable. We conclude that aPIC and aPIC+SA may be useful in terms of yielding relatively small medians with small variances.
Table 3: The results for Case 3 and Case 4.

| Case 3 | Case 4 |
|--------|--------|
|        | $n = 20$ | $n = 50$ | $n = 50$ | $n = 100$ | $n = 200$ |
|        | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC   | 1.00 | 0.00 | 0.10 | 0.51 | 4.31 | 1.00 | 0.00 | 0.45 | 7.44 | 30.93 |
| aPIC+SA| 0.12 | 0.90 | 0.80 | 0.46 | 4.32 | 0.60 | 0.82 | 0.71 | 8.86 | 32.20 |
| DIC    | 1.00 | 0.00 | 0.10 | 1.17 | 4.63 | 1.00 | 0.00 | 0.43 | 9.41 | 31.79 |
| DIC+SA | 0.37 | 0.71 | 0.66 | 1.12 | 4.61 | 0.73 | 0.57 | 0.63 | 9.94 | 32.35 |
| Blasso | 1.00 | 0.00 | 0.10 | 0.30 | 4.23 | 1.00 | 0.00 | 0.44 | 8.02 | 31.41 |
| Blasso+SA | 0.02 | 0.99 | 0.87 | 0.26 | 4.20 | 0.61 | 0.75 | 0.67 | 10.51 | 33.78 |
| WAIC   | 1.00 | 0.00 | 0.10 | 1.38 | 4.73 | 1.00 | 0.00 | 0.46 | 4.29 | 28.25 |
| WAIC+SA| 0.58 | 0.55 | 0.57 | 1.32 | 4.68 | 0.84 | 0.49 | 0.69 | 4.28 | 28.29 |
| Lasso  | 0.77 | 0.39 | 0.43 | 1.14 | 4.70 | 0.71 | 0.65 | 0.67 | 7.94 | 31.05 |

Table 4: The numbers of observations and predictors for real datasets.

| # of observations | diabetes | Boston housing | Parkinson | communities and crimes |
|-------------------|----------|----------------|-----------|------------------------|
|                   | 442      | 506            | 5875      | 2195                   |
| # of predictors   | 10       | 13             | 19        | 102                    |
Fig. 1: Boxplots of the PSE. (a) shows the result for the diabetes, (b) that for the Boston housing, (c) that for the Parkinson, (d) that for the communities and crimes.

5. Concluding remarks

We proposed the information criterion aPIC, which is obtained by evaluating the Bayesian predictive distribution for Bayesian lasso regression, for the selection of appropriate values of hyperparameters included in a prior distribution. For derivations of this criterion, we need to approximate the Laplace prior distribution for the coefficients. The prior distribution was approximated by the normal distribution from the viewpoints of minimizing the Kullback-Leibler information between the Laplace distribution and normal distributions. Numerical examples showed that our proposed procedure is useful in terms of prediction, estimation, and model selection accuracies.
Model Selection Criteria for Bayesian Lasso

It is important to introduce information criteria by evaluating the Bayesian predictive distribution for logistic, Poisson, and Cox regressions estimated by the Bayesian lasso. In addition, it would also be interesting to derive model selection criteria for other sparse regularization methods from Bayesian viewpoints, e.g., elastic net, SCAD, and MCP by a Bayesian approach. We leave these topics as future research.

Acknowledgments

The authors would like to thank the anonymous reviewers for helpful comments and suggestions. This work was partially supported by a Grant-in-Aid for Young Scientists (B) (#24700280, #15K15947) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

REFERENCES

Andrews, D. F. and Mallows, C. L. (1974). Scale mixture of normal distributions. *Journal of the Royal Statistical Society Series B* 36, 99–102.
Biswas, S. and Lin, S. (2012). Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* 68, 587–597.
Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 127–135.
Brady, A. R., Thompson, S. G., Fowkes, F. G. R., Greenhalgh, R. M. and Powell, J. T. (2004). Abdominal aortic aneurysm expansion risk factors and time intervals for surveillance. *Circulation* 110, 16–21.
Dicker, L., Huang, B. and Lin, X. (2013). Variable selection and estimation with the seamless-$l_0$ penalty. *Statistica Sinica* 23, 929–962.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *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., Hastie, T. and Tibshirani, R. (2010). Regularization paths fo generalized linear models via coordinate descent. *Journal of Statistical Software* 33, 1–22.
Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* 7, 457–511.
Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian and New Zealand Journal of Statistics* 53, 423–442.
Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing* 20, 221–229.
Hoshina, I. (2012). Sparse regression modeling via the Bayesian lasso (in Japanese). *Bulletin of the Computational Statistics of Japan* 25, 73–85.
Kabán, A. (2007). On Bayesian classification with Laplace priors. *Pattern Recognition Letters* 28, 1271–1282.
Kim, D., Kawano, S. and Konishi, S. (2012). Predictive information criteria for Bayesian nonlinear regression models. *Bulletin of Informatics and Cybernetics* 44, 17–28.
Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics – Theory and Methods* 26, 2223–2246.
Konishi, S. and Kitagawa, G. (2008). Information Criteria and Statistical Modeling. New York: Springer.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis 5, 369–412.

Leng, C., Tran, M.-N. and Nott, D. J. (2014). Bayesian adaptive lasso. Annals of the Institute of Statistical Mathematics 66, 221–244.

Li, J., Das, K., Fu, G., Li, R. and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. Bioinformatics 27, 516–523.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. Bayesian Analysis 5, 151–170.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8, 1235–1246.

Mutshida, C. M. and Sillanpää, M. J. (2010). Extended Bayesian lasso for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186, 1067–1075.

O’Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. Bayesian Analysis 4, 85–118.

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

Pieters, R. and Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. Journal of Marketing 68, 36–50.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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 Series B 64, 583–639.

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

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

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. 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 Series B 67, 301–320.

(Received: January 30, 2015, Accepted: June 15, 2015)