Neural Networks 61 (2015) 22–31

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

An efficient sampling algorithm with adaptations for Bayesian variable selection

Takamitsu Araki [a,*], Kazushi Ikeda [b], Shotaro Akaho [a]

[a] Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology, Japan

[b] Graduate School of Information Science, Nara Institute of Science and Technology, Japan

ARTICLE INFO

Article history:
Received 10 May 2014
Received in revised form 8 September 2014
Accepted 26 September 2014
Available online 7 October 2014

Keywords:
Indicator model selection
Gibbs variable selection
Kuo and Mallick's method
Adaptive Markov chain Monte Carlo
Convergence
Bayesian logistic regression model

ABSTRACT

In Bayesian variable selection, indicator model selection (IMS) is a class of well-known sampling algorithms, which has been used in various models. The IMS is a class of methods that uses pseudo-priors and it contains specific methods such as Gibbs variable selection (GVS) and Kuo and Mallick's (KM) method. However, the efficiency of the IMS strongly depends on the parameters of a proposal distribution and the pseudo-priors. Specifically, the GVS determines their parameters based on a pilot run for a full model and the KM method sets their parameters as those of priors, which often leads to slow mixings of them. In this paper, we propose an algorithm that adapts the parameters of the IMS during running. The parameters obtained on the fly provide an appropriate proposal distribution and pseudo-priors, which improve the mixing of the algorithm. We also prove the convergence theorem of the proposed algorithm, and confirm that the algorithm is more efficient than the conventional algorithms by experiments of the Bayesian variable selection.

1. Introduction

Bayesian variable selection plays an important role in causal analysis, prediction and classification. It calculates a posterior distribution of coefficients and inclusion of covariates in the statistical models. The posterior distribution is mainly used in the two Bayesian inferences. One is extraction of the important covariates, and the other is construction of a predictive distribution which is a powerful tool for various purposes such as pattern recognition and prediction. Since the posterior distribution cannot be obtained in a closed form, Markov chain Monte Carlo (MCMC) methods are applied to the estimation. The MCMC methods can efficiently generate samples from such a complex distribution by simulating a Markov chain that converges to the target distribution and are often used to calculate the statistics of the posterior distribution in the Bayesian analysis (Robert & Casella, 2004). However, the standard MCMC methods cannot efficiently generate samples from the posterior distribution in the Bayesian variable selection due to its inherent complex structure.

In order to generate samples from the posterior efficiently, indicator model selection (IMS; Dellaportas, Forster, & Ntzoufras, 2002 and Kuo & Mallick, 1998) has been proposed. The IMS is a class of the MCMC methods which uses pseudo-priors and contains Gibbs variable selection (GVS; Dellaportas et al., 2002) and Kuo and Mallick's (KM) method (Kuo & Mallick, 1998). For conditionally non-conjugate models in particular, the IMS is more efficient than the other algorithms based on the MCMC methods (O'Hara & Sillanpaa, 2009) such as Stochastic Search Variable Selection (SSVS; George & McCullogh, 1993). The conditionally non-conjugate models such as generalized linear models and nonlinear models have the conditional posterior densities that cannot be obtained in the closed form. Reversible jump MCMC (RJMCMC; Green, 1995), which is the well-known MCMC method for Bayesian model selection, is also applicable to the posterior of the Bayesian variable selection for the both conditionally conjugate and non-conjugate models, but is less efficient than the GVS, Dellaportas et al. (2002). From those reasons, the IMS seems to be widely used to the Bayesian variable selection in various models.

The IMS introduces the pseudo-priors to improve sampling efficiency, but each algorithm included in the IMS hardly obtains the appropriate pseudo-priors as follows. In the GVS, the parameters of the pseudo-priors are determined by a pilot run for a full model (Dellaportas et al., 2002). However, the pseudo-priors with the obtained parameters have different distribution properties, a position

* Correspondence to: Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan. Tel.: +81 29 861 9488; fax: +81 29 861 6636.

E-mail addresses: tk-araki@aist.go.jp (T. Araki), kazushi@is.naist.jp (K. Ikeda), s.akaho@aist.go.jp (S. Akaho).

http://dx.doi.org/10.1016/j.neunet.2014.09.010
0893-6080/

and a shape, from the marginal posteriors of coefficients due to the correlation of the posterior distribution of coefficient parameters, which causes a slow mixing of the GVS. In the KM method, the pseudo-priors are the same distributions as the priors of the coefficients. When the priors are vague or non-informative priors, e.g., the Jeffreys prior, the pseudo-priors have also the different distribution properties from the marginal posteriors of coefficients.

The IMS typically uses a Metropolis algorithm as a sampling method of the coefficient parameters for the conditionally non-conjugate models. The covariance matrix of the proposal distribution is also calculated by the samples from the pilot run for the full model (Paroli & Spezia, 2007). However, since the samples from the full model do not have enough information to estimate an appropriate scale of the proposal distribution, the proposal covariances above often lead to a slow convergence of the IMS.

The IMS can be regarded as a special case of auxiliary variable methods (AVMs). The AVMs are the advanced MCMC methods that use auxiliary distributions and contain also Parallel Tempering and cluster Monte Carlo method and so on. The AVMs were extended to an adaptive MCMC for AVMs that adapts the parameters of the AVMs while it runs (Araki & Ikeda, 2013), and the convergence of the algorithm was proved.

To solve the parameter setting problems of the IMS, we propose an adaptive IMS that adapts the proposal covariances and pseudo-priors by learning the means and the covariances of the coefficient posterior and the scale of the proposal distribution while it runs. We prove the convergence of the proposed algorithm by using the convergence theorem of the adaptive MCMC for AVMs in Araki and Ikeda (2013), since our algorithm can be formulated as the adaptive MCMC for AVMs. We also show that our algorithm can obtain appropriate parameters during its run and is more efficient than the conventional algorithms through their applications to the Bayesian variable selection of a linear regression model and logistic regression models.

The rest of this paper is organized as follows. The conventional IMS is described in Section 2. The adaptive IMS is proposed in Section 3 and its convergence theorem is proved in Section 4. In Section 5, performance of the proposed algorithm is evaluated through numerical experiments. Finally, we give conclusions in Section 6.

2. Indicator model selection

We consider the following formulation of statistical models as the models that apply the Bayesian variable selection in this paper. The p-variate statistical models have the coefficient parameter vector, $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$, associated with covariates $x_j, j = 1, \ldots, p$. For example, a regression model, one of the simplest multivariate models, is written as $y = \sum_{j=1}^{p} x_j\theta_j + \epsilon$, where y is the response to x and ϵ is a noise.

The IMS sets $\theta_j = \gamma_j\beta_j, j = 1, \ldots, p$, where $\beta = (\beta_1, \ldots, \beta_p) \in \Theta$, and $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$ is an indicator variable vector that represents which covariates are included in the model. That is, γ_j takes one if the covariate x_j is included, and zero otherwise. In order to generate samples from the posterior of γ, θ efficiently, the IMS generates the samples from the posterior of γ, β. If γ_j takes one, β_j is equal to the coefficient θ_j; otherwise β_j is distributed according to the pseudo-prior $f_{\lambda_j}(\beta_j)$, where λ_j is two dimensional value denoting the mean and the variance. The pseudo-priors are not included in the posterior of θ and γ, but facilitate to produce the sample sequence from the posterior of γ and β. The prior of β_j given γ_j is

$$f_{\lambda_j}(\beta_j|\gamma_j) = \gamma_j f_j(\beta_j) + (1 - \gamma_j) f_{\lambda_j}(\beta_j), \tag{1}$$

where $f_j(\beta_j)$ is a j-th coefficient prior.

The IMS conducts Gibbs sampling steps for γ and β_γ, and a Metropolis–Hastings step for $\beta_{\setminus\gamma}$ by turns, where β_γ denotes the components of β included in the model, whose corresponding indicators, γ_j, take one, and $\beta_{\setminus\gamma}$ consists of the others. The Gibbs sampling step produces samples from the conditional posterior distributions

$$f_\lambda(\gamma_j|\gamma_{-j}, \beta, D) \propto f(D|\beta, \gamma) \prod_{k=1}^{p} f_{\lambda_k}(\beta_k|\gamma_k) f_k(\gamma_k), \quad j = 1, \ldots, p,$$

$$f_\lambda(\beta_{\setminus\gamma}|\gamma, \beta_\gamma, D) = \prod_{\beta_j \in \beta_{\setminus\gamma}} f_{\lambda_j}(\beta_j),$$

where $\lambda = (\lambda_1, \ldots, \lambda_p) \in \Lambda \subset \mathbb{R}^{2p}$, and γ_{-j} denotes the entries of γ except γ_j, $f(D|\beta, \gamma)$ is the likelihood of the observation data D with β and γ, and $f_k(\gamma_k)$ is the prior of γ_k. The Metropolis–Hastings step executes the Metropolis–Hastings update for

$$f(\beta_\gamma|\gamma, \beta_{\setminus\gamma}, D) \propto f(D|\beta, \gamma) \prod_{\beta_j \in \beta_\gamma} f_j(\beta_j). \tag{2}$$

From the practical viewpoint, if this conditional distribution (2) can be obtained analytically, this step directly samples from the distribution, that is, the Gibbs sampling is conducted; otherwise the Metropolis sampling is applied. In this paper, we employ the Metropolis sampling because it can be applied to more various models.

The GVS is a specific case of the IMS, in which the parameters of the pseudo-priors are estimated by the samples from the full model. On the other hand, the KM is a case in which the pseudo-priors are the same distribution as the coefficient priors.

The pseudo-priors should well approximate the marginal coefficient posteriors, $f(\beta_j|\gamma_j = 1, D)$, for the sake of a well mixing of the IMS. The pilot run in the GVS samples from the posterior of the coefficients of the full model, $f(\beta|\gamma_1 = \cdots = \gamma_p = 1, D)$, and the means and the covariances of the coefficient posterior, $\mu^* \in U \subseteq \Theta$ and $\Sigma^* \in \Xi$, are estimated by the sample means, $\hat{\mu}$, and the sample covariances, $\hat{\Sigma}$, respectively, where Ξ denotes a subset of $p \times p$ real positive definite symmetric matrices. The GVS employs the estimated parameters as those of the pseudo-priors (Dellaportas et al., 2002; Paroli & Spezia, 2007). However, the distribution properties, the shape and the position, of the pseudo-priors with the means $\hat{\mu}_j$ and variances $\hat{\Sigma}_{jj}$ are different from those of the marginal coefficient posteriors, $f(\beta_j|\gamma_j = 1, D)$, because the distribution properties of the marginal coefficient posteriors are different from those of the marginal posterior of the full model from which the pilot run generates samples. The pseudo-priors in the KM method have also different distribution properties from the marginal coefficient posteriors, since the pseudo-priors correspond to the coefficient priors.

The proposal distribution should provide an appropriate mean Metropolis-acceptance rate, typically 0.234 in multidimensional settings (Roberts, Gelman, & Gilks, 1997), for the sake of rapid mixing of the IMS. The IMS employs the covariances estimated by the samples from the posterior of the full model as those of the proposal distribution (Dellaportas et al., 2002; Paroli & Spezia, 2007). The covariances of the proposal distribution is typically set as $c\,\hat{\Sigma}_{ij}$, where $c \in \mathcal{C} \subseteq \mathbb{R}_+$ is a scale parameter. Roberts et al. (1997) theoretically showed that if the model is high dimensional the scale parameter $c = (2.38)^2/p$ is appropriate, and the value has been used as a standard value. However, such proposal distribution often brings about the inappropriate mean acceptance rate, because the scale of the proposal distribution that leads to the appropriate mean acceptance rate practically depends on the dimension and the shape of the coefficient posterior.

3. Adaptive indicator model selection

We propose the adaptive IMS that adapts the pseudo-prior parameters and the proposal covariances by learning the covariances and the means of the coefficient posterior and the scale parameter while the IMS is running. The two learning algorithms are described below.

Since the correlation of the proposal distribution, the variances and the means of the pseudo-priors should correspond with those of the coefficient posterior, the covariances and the means of the coefficient posterior, Σ^* and μ^*, are necessary. To obtain Σ^* and μ^*, the covariance parameters Σ and the mean parameters μ are updated by using only the samples $\beta_j^{(n)}$ from the coefficient posterior, $f(\beta_j|D, \gamma_j = 1)$. Thus, the updates use the number of sampling β_j from the coefficient posterior, $a_j^{(n)}$, but not the iteration number n. The update equations are as follows.

$$\begin{aligned}
\mu_j^{(n+1)} &\leftarrow \mu_j^{(n)} + \gamma_j^{(n+1)} h(a_j^{(n)})(\beta_j^{(n+1)} - \mu_j^{(n)}),\\
\Sigma_{ij}^{(n+1)} &\leftarrow \Sigma_{ij}^{(n)} + \gamma_j^{(n+1)}\gamma_i^{(n+1)} u(a_i^{(n)}, a_j^{(n)})\\
&\quad \times \left((\beta_j^{(n+1)} - \mu_j^{(n)})(\beta_i^{(n+1)} - \mu_i^{(n)}) - \Sigma_{ij}^{(n)}\right), \qquad (3)\\
a_j^{(n+1)} &\leftarrow a_j^{(n)} + \gamma_j^{(n+1)}, \quad j = 1, \ldots, p,\ i = 1, \ldots, p,
\end{aligned}$$

where $a_j^{(0)} = 1$, and $a_j^{(n)}$ is incremented when β_j is sampled from the coefficient posterior, $f(\beta_j|D, \gamma_j = 1)$. The learning coefficients $h(n)$ and $u(n, m)$ are decreasing functions that satisfy $\lim_{n\to\infty} h(n) = 0$ and $\lim_{n\to\infty, m\to\infty} u(n, m) = 0$, respectively. Note that μ_j and Σ_{ij} are updated by only the samples $\beta_j^{(n)}$ from the coefficient posterior, $f(\beta_j|D, \gamma_j = 1)$, and the samples $\beta_i^{(n)}$, $\beta_j^{(n)}$ from $f(\beta_i, \beta_j|D, \gamma_i = 1, \gamma_j = 1)$, respectively.

The proposal distribution should lead to the appropriate mean Metropolis acceptance rate, $\alpha \in (0, 1)$, 0.234 in many cases. To achieve the rate α, the scale parameter c is updated as

$$c^{(n+1)} \leftarrow c^{(n)} + s(n)(ER^{(n+1)} - \alpha), \qquad (4)$$

where $ER^{(n+1)}$ is a variable that takes one if the proposal value in the Metropolis sampling of $\beta_{\gamma^{(n+1)}}$ is accepted at time n, and zero otherwise. The learning coefficient, $s(n)$, is a decreasing function of n that satisfies $\lim_{n\to\infty} s(n) = 0$.

The learned parameters $\mu^{(n)}$, $\Sigma^{(n)}$, and $c^{(n)}$ are introduced to the pseudo-priors and the proposal distribution. That is, the pseudo-priors have the mean $\mu_j^{(n)}$ and the variance $\Sigma_{jj}^{(n)}$ and the covariance of the proposal distribution are $c^{(n)}\Sigma_{ij}^{(n)}$ at the $(n+1)$-th iteration.

A pseudo code of the adaptive IMS is given in Algorithm 1.

Numerically, the appropriate convergence order of the learning coefficients, $h(n)$ and $s(n)$, is $O(1/n)$, and that of $u(n, m)$ is $O(1/\sqrt{nm})$.

4. Convergence theorem

We prove the convergence of the adaptive IMS by applying the theorem of Araki and Ikeda (2013) which assures the convergence of the adaptive MCMC for AVMs, since the adaptive IMS can be formulated as the adaptive MCMC for AVMs (See Appendix A for details).

The target distribution of the adaptive IMS, formulated as the adaptive MCMC for AVMs, is the posterior distribution of the indicator variables γ_j and coefficients $\theta_j = (\gamma_j\beta_j)$, $f(\theta, \gamma|D)$. Here the convergence of the adaptive IMS means that the adaptive IMS is ergodic, that is, the samples $\gamma_j^{(n)}$ and $\theta_j^{(n)}$ generated by the adaptive IMS converge in distribution to the posterior distribution $f(\theta, \gamma|D)$. (The exact definition of the ergodicity is in Appendix B.) The convergence is assured as follows.

Algorithm 1 Adaptive IMS algorithm

Initialize $\beta^{(0)}$, $\gamma^{(0)}$, $\Sigma^{(0)}$, $\mu^{(0)}$ and $c^{(0)}$.
for $n = 0$ to $N - 1$ **do**
  (*Gibbs sampling step*)
  for $j = 1$ to p **do**
    $\gamma_j^{(n+1)} \sim f_{\lambda_j^{(n)}}(\gamma_j|\gamma_{-j}^{(n)}, \beta^{(n)}, D)$, where $\lambda_j^{(n)} = (\mu_j^{(n)}, \Sigma_{j,j}^{(n)})$ and $\gamma_{-j}^{(n)} = (\gamma_1^{(n+1)}, \ldots, \gamma_{j-1}^{(n+1)}, \gamma_{j+1}^{(n)}, \ldots, \gamma_p^{(n)})$.
  end for
  $\beta_{\setminus\gamma^{(n+1)}}^{(n+1)} \sim \prod_{\beta_j\in\beta_{\setminus\gamma^{(n+1)}}} f_{\lambda_j^{(n)}}(\beta_j)$.
  (*Metropolis sampling step*)
  Generate $\beta_{\gamma^{(n+1)}}^{(n+1)}$ via the Metropolis algorithm for $f(\beta_{\gamma^{(n+1)}}|\gamma^{(n+1)}, \beta_{\setminus\gamma^{(n+1)}}^{(n+1)}, D)$, which has the proposal covariance matrix $c^{(n)}\Sigma_{\gamma^{(n+1)}}^{(n)}$, where Σ_γ denotes the covariance matrix that consists of the covariances Σ_{ij}, where $\gamma_i = 1$ and $\gamma_j = 1$.
  (*Parameter learning step*)
  Update $(\mu^{(n)}, \Sigma^{(n)})$ to $(\mu^{(n+1)}, \Sigma^{(n+1)})$ by the Eq. (3).
  Update $c^{(n)}$ to $c^{(n+1)}$ by Eq. (4).
end for

Theorem 1. *The adaptive IMS is ergodic, if the following conditions hold:*

(s1) *The support $S \times \Gamma$ of the posterior distribution of β and γ, $f_\lambda(\beta, \gamma|D)$, is compact for $\lambda \in \Lambda$, where $S \subseteq \Theta$, $\Gamma = \{0, 1\}^p$ and $\Lambda = \mathcal{U} \times [\varsigma, \varsigma']^p$, where $\mathcal{U}$ is a bounded set on $\mathbb{R}^p$ and $\varsigma > 0$. For $\gamma \in \Gamma$, $g_\gamma(\lambda, \beta) \equiv f_\lambda(\beta, \gamma|D)$ is continuous and positive on $\Lambda \times S$.*

(s2) *The family of proposal densities $\{q_{(\Sigma,c)}\}_{(\Sigma,c)\in\Xi\times\mathcal{C}}$ is continuous on $\Theta \times \Theta \times \Xi \times \mathcal{C}$ and positive on $S^2 \times \Xi \times \mathcal{C}$, where $\mathcal{C}$ is a bounded set on $\mathbb{R}_+$, $S^2 = S \times S$ and $\Xi = \{\Sigma|\varsigma I_p \le \Sigma \le \varsigma' I_p\}$, where I_p denotes a p-dimensional identity matrix.*

Proof. See Appendix C. □

5. Numerical validations

To compare numerically the convergence rates and the efficiencies of the adaptive IMS and the conventional algorithms, these algorithms were applied to the Bayesian variable selection for two models. First, the Bayesian variable selection was applied to the normal linear regression model as the simplest example. For linear models, the performances among different methods are not much different in theory. Second, the logistic regression model was used because it is conditionally non-conjugate and the adaptive IMS is expected to improve the efficiency of sampling for the Bayesian variable selection of the conditionally non-conjugate model.

The conventional algorithms used in the numerical experiments are the GVS, the KM method, the SSVS and the RJMCMC. The SSVS induces a spike prior, whose probability concentrates in neighborhood of 0. When $\gamma_j = 0$, the prior of the coefficient β_j is the spike prior. The RJMCMC in the Bayesian variable selection, randomly chooses a covariate, and inverts a value of the corresponding indicator variable, and deletes or regenerates the corresponding coefficient parameter. All the conventional algorithms described above use the Metropolis sampling, and the covariances of the proposal distribution are estimated by the samples from the full model, which is the same way as the IMS.

Throughout the numerical experiments, we consistently used the setting for the algorithms described below.

In our algorithm, the initial parameter values $\mu^{(0)}$, $\Sigma^{(0)}$ and $c^{(0)}$ were set as follows. $\mu^{(0)}$ was a mode of the full model posterior density of β, denoted by $\hat{\beta}$. $\Sigma^{(0)}$ was the inverse of the negative

Table 1
The learning coefficients and the target mean acceptance rate.

$h(n)$	$u(n, m)$	$s(n)$	α
$1/(n+50)$	$1/\sqrt{(n+50)(m+50)}$	$1/(n+500)$	0.234

Table 2
The coefficients, θ_j^*, and the covariances, σ_{ij}^*, of the true distribution of the covariates.

i, j	1	⋯	5	⋯	31	⋯	35	⋯	51	⋯	55	⋯
θ_j^*	-0.5			0	-0.1			0	1			0
σ_{ij}^*	0.8				0							

i, j	⋯	71	⋯	75	⋯	96	⋯	100
θ_j^*	0	-0.5		0		0.1		
σ_{ij}^*	0	0.7						

Hessian of the log full model posterior density at the mode $\hat{\beta}$, and $c^{(0)} = (2.38)^2/p$. The learning coefficients and the target mean acceptance rate are described in Table 1.

In the conventional algorithms, the proposal scale was $c = (2.38)^2/p$. The initial sample values of the pilot runs were $\hat{\beta}$.

The adaptive IMS and all the conventional algorithms were given the same initial sample values $\beta^{(0)}$ and $\gamma^{(0)}$ and the same burn-in period. In the adaptive IMS, the GVS and the KM method, the sample sets used in estimation were chosen from every 10 samples.

We also compare the performance of the Bayesian variable selection using the MCMC method to that of the non-Bayesian variable selection, the stepwise forward/backward feature selection using Akaike's Information Criterion (SFS; Hastie & Pregibon, 1992) and sparse estimator via L_1-penalized likelihood called Lasso estimator (Tibshirani, 1996) with 10-fold cross validation.

5.1. Normal linear regression model

We estimated the marginal probability of inclusion for each of p covariates, x_j, and the predictive distribution for the normal linear regression model using the synthetic data.

The normal linear regression model is,

$$f(y|x, \beta, \gamma, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \sum_{j=1}^{p} x_j \beta_j \gamma_j)^2}{2\sigma^2} \right),$$

where $y \in \mathbb{R}$ is a response variable. The coefficient priors and the priors of γ_j and the pseudo-priors are

$$f_j(\beta_j) = N(\beta_j|\mu_{\beta_j}, \sigma_{\beta_j}^2),$$
$$f_j(\gamma_j) = \tau_j^{\gamma_j}(1-\tau_j)^{1-\gamma_j},$$
$$f_{\lambda_j}(\beta_j) = N(\beta_j|\tilde{\mu}_j, \tilde{\sigma}_j^2), \quad \lambda_j = (\tilde{\mu}_j, \tilde{\sigma}_j^2),$$

where $N(\cdot|\mu, \sigma^2)$ is a Gaussian density with mean μ and variance σ^2, and $0 < \tau_j < 1$. The prior of the noise variance σ^2 is

$$f(\sigma^2) = IG(\sigma^2|a_{\sigma^2}, b_{\sigma^2}),$$

where $IG(\cdot|a, b)$ is an inverse-gamma density with shape a and rate b. The hyper-parameters were $\mu_{\beta_j} = 0$, $\sigma_{\beta_j}^2 = 10^2$, $\tau_j = 0.5$, $a_{\sigma^2} = 0.1$ and $b_{\sigma^2} = 0.1$.

The synthetic data of size 300 were independently identically distributed according to the normal linear regression model, which has $p = 100$ covariates and the coefficients θ^* (Table 2). The coefficients θ^* imply that the normal linear regression model depends on the only 25 covariates. The covariates were generated from the normal distribution with mean 0 and variance 1 and covariances $\sigma_{ij}^* = 0.8$ for $i, j = 1, \ldots, 30$, $\sigma_{ij}^* = 0.7$ for $i, j = 71, \ldots, 100$, and $\sigma_{ij}^* = 0$ otherwise (Table 2).

We generated samples from the posterior given the synthetic data and calculated the estimated marginal probabilities of inclusions $\hat{P}_j^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \gamma_j^{(i)}$ and their mean absolute errors (Atchade, 2011)

$$\mathrm{EP}^{(n)} = \frac{1}{p}\sum_{j=1}^{p} |\hat{P}_j^{(n)} - P_j^*|, \tag{5}$$

where $P_j^* = 1$ if x_j is included in the true model, and $P_j^* = 0$ otherwise. In order to evaluate the properties of not only the samples $\gamma^{(n)}$ but also $\beta^{(n)}$, we also computed the estimated cross entropies between the estimated predictive distributions, $\tilde{f}^{(n)}(y|x) = \frac{1}{n}\sum_{i=1}^{n} f(y|x, \beta^{(i)}, \gamma^{(i)}, \sigma^{2(i)})$, and the true model,

$$\mathrm{CE}^{(n)} = -\frac{1}{10^3}\sum_{i=1}^{10^3} \log \tilde{f}^{(n)}(y^{(i)}|x^{(i)}), \tag{6}$$

where $x^{(i)}, y^{(i)}$ were generated from the true model independently of the training data.

We ran the adaptive IMS, the GVS, the KM method and the SSVS for 3×10^4 iterations 10 times independently. The RJMCMC was run for 99×10^4 iterations so that the computational times of the algorithms roughly correspond. We choose the samples of the RJMCMC every 330 iterations. Since the normal linear model is conditionally conjugate, all the algorithms do not use the Metropolis sampling and only the GVS conducts the pilot run to estimate the pseudo-priors. The GVS conducted its pilot runs for 50 and 100 iterations. The initial sample values were $\beta_j^{(0)} \sim U(-1, 1)$, $\gamma_j^{(0)} \sim Be(0.5)$ and $\sigma^{2(0)} \sim U(0, 10)$ independently, where $U(a, b)$ is the uniform distribution of the interval (a, b) and $Be(a)$ is the Bernoulli distribution with success probability a.

$\mathrm{EP}^{(n)}$ and $\mathrm{CE}^{(n)}$ of the algorithms except the RJMCMC converged at almost the same rates due to the following reasons (Fig. 1). The normal linear regression model is simple and conditionally conjugate, where the parameters can be directly sampled from the conditional distribution. The parameters that need to be tuned are only those of the pseudo-priors. Therefore, we proposed our algorithm for nonlinear models, which are conditionally non-conjugate. The RJMCMC was much slow, because it uses the general approach for the Bayesian model selection and may need the proposal distribution fitted by hand turning. $\mathrm{EP}^{(n)}$ and $\mathrm{CE}^{(n)}$ of the SSVS converged to different values from those of the adaptive IMS, the GVS and the KM method (Fig. 1), because it samples from the approximate distribution of the posterior.

We calculated the mean absolute error of variable selection EP and the cross entropy CE of the models estimated by the SFS and the Lasso. The mean absolute error is $\mathrm{EP} = \frac{1}{p}\sum_{j=1}^{p} |\hat{P}_j - P_j^*|$, where $\hat{P}_j = 1$ if the corresponding covariate is included in the estimated model, and $\hat{P}_j = 0$ otherwise. The estimated cross entropy between the true model and the estimated model, $f(y|x, \hat{\theta}, \hat{\sigma}^2)$, where $\hat{\theta}$ and $\hat{\sigma}^2$ are estimated parameters, is $\mathrm{CE} = -\frac{1}{10^3}\sum_{i=1}^{10^3} \log \tilde{f}^{(n)}(y^{(i)}|x^{(i)})$, where $x^{(i)}, y^{(i)}$ were generated from the true model as described above.

Both the mean absolute error and the cross entropy for the Bayesian variable selection based on our algorithm (Eqs. (5) and (6)) were lower than those for the SFS and the Lasso (Table 3).

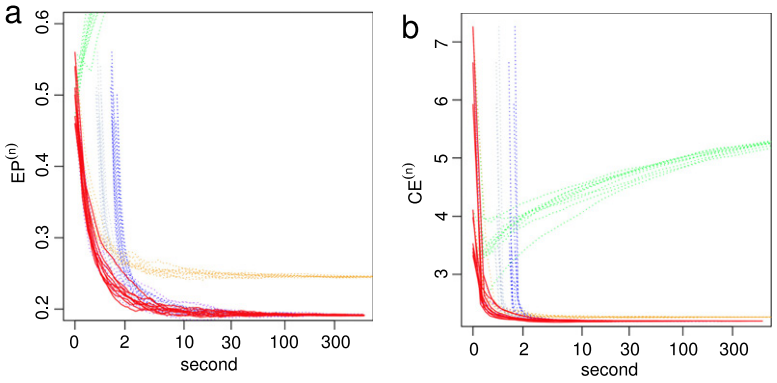

Fig. 1. Trace plots of the $EP^{(n)}$ (a) and $CE^{(n)}$ (b) by the adaptive IMS (red line), the GVS with the pilot run of 50 iterations (gray dotted line) and 100 iterations (blue dotted line), the KM method (purple dotted line), the SSVS (marigold dotted line) and the RJMCMC (green dotted line) in 10 runs. The $EP^{(n)}$ and $CE^{(n)}$ by the GVS are plotted after its pilot run completes. The computational time is transformed logarithmically as $\log_{10}(\text{time} + 1)$. Note that almost of the plot curves by the KM method are identical to those of the adaptive IMS and overlapping together. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
The mean absolute errors and the cross entropies for the Lasso, the SFS and the Bayesian variable selection using the samples generated by the adaptive IMS. The values for the Bayesian variable selection were $EP^{(N)}$ and $CE^{(N)}$ obtained by the adaptive IMS, which were plotted in Fig. 1. Each value for the Bayesian variable selection is the worst value in the 10 runs of our algorithm.

	Mean absolute error	Cross entropy
Bayesian variable selection	0.192	2.195
Lasso	0.240	2.200
SFS	0.270	2.348

5.2. Logistic regression model

The Bayesian variable selection for the logistic regression model was conducted for synthetic data and cardiac arrhythmia data. Throughout the two numerical experiments, we consistently used the setting for the logistic regression model described below.

The logistic regression model is,

$$f(y|x, \beta, \gamma) = \frac{1}{1 + \exp\left(-y\left(\sum_{j=1}^{p} x_j \beta_j \gamma_j\right)\right)},$$

where $y \in \{-1, 1\}$ is a response variable. The priors and pseudo-priors are

$$f_j(\beta_j) = N(\beta_j|\mu_{\beta_j}, \sigma^2_{\beta_j}),$$
$$f_j(\gamma_j) = \tau_j^{\gamma_j}(1 - \tau_j)^{1-\gamma_j},$$
$$f_{\lambda_j}(\beta_j) = N(\beta_j|\mu_j, \sigma_j^2), \quad \lambda_j = (\mu_j, \sigma_j^2).$$

The hyper-parameters were $\mu_{\beta_j} = 0$, $\sigma^2_{\beta_j} = 9$ and $\tau_j = 0.5$.

5.2.1. Synthetic data

We evaluated $EP^{(n)}$ and $CE^{(n)}$ defined in Eqs. (5) and (6) for the logistic regression using the synthetic data. We also compare these criteria for the Bayesian variable selection using our algorithm to those for the Lasso and the SFS.

The synthetic data of size 10^3 were independently identically distributed to the logistic regression model, which has the same covariates and the coefficients θ^* as those of the normal linear model above (Table 2).

The adaptive IMS, the GVS and the KM method were run for 2×10^5 iterations, and the SSVS and the RJMCMC were run for 10^6 and 66×10^5 iterations, respectively. Each of these algorithms was run 10 times independently. The algorithms except the adaptive IMS conducted its pilot runs for 10^3, 2×10^3, 5×10^3 and 10^4 iterations. The initial sample values were independently randomly chosen as $\beta_j^{(0)} \sim U(-2 + \hat{\beta}_j, 2 + \hat{\beta}_j)$ and $\gamma_j^{(0)} \sim Be(0.5)$.

$EP^{(n)}$ by the adaptive IMS converged fastest of those by all the algorithms for all pilot runs, and the variance of the errors by the adaptive IMS was the smallest (Fig. 2). $CE^{(n)}$ of the adaptive IMS also converged fastest and stablest of those of the other algorithms for all pilot runs (Fig. 3). The convergence rates and the stabilities of the conventional algorithms were inferior to those of the adaptive IMS regardless of the iteration number of their pilot runs.

The learning parameters $\mu_2^{(n)}$, $\Sigma_{22}^{(n)}$ and $c^{(n)}$ converged quickly and stably (Fig. 4), and the others also converged as fast and stable as them.

The mean acceptance rates in the adaptive IMS were close to desirable value $\alpha = 0.234$, but those in the GVS and the KM method were not (Table 4). This leads to well mixing of our algorithm and causes a slow convergence rate of the GVS and the KM method.

To compare the learning accuracies of the variances and the means of the posterior of θ obtained by our algorithm and the pilot run of the GVS, we calculated their mean absolute errors

$$E_\mu = \frac{1}{p}\sum_{j=1}^{p} |\tilde{\mu}_j - \dot{\mu}_j| \quad \text{and} \quad E_\Sigma = \frac{1}{p}\sum_{j=1}^{p} |\tilde{\Sigma}_{jj} - \dot{\Sigma}_{jj}|,$$

where $\tilde{\mu}_j$ and $\tilde{\Sigma}_{jj}$ are $\mu_j^{(N)}$ and $\Sigma_{jj}^{(N)}$ in the adaptive IMS, and $\hat{\mu}_j$ and $\hat{\Sigma}_{jj}$ in the GVS, which are the mean and the variance of the pilot run samples. The target parameters $\dot{\mu}_j$ and $\dot{\Sigma}_{jj}$ represent the true means and variances of the posterior of coefficients, but were practically estimated by averaging the 10 estimates of the coefficient's posterior means and variances. The estimates were computed from the samples from the GVS for 10^6 iterations with the pilot run of 5×10^5 iterations and with the tuned scale parameter in which the mean acceptance rates were in (0.238, 0.241).

The mean and variance parameters learned by the adaptive IMS were much closer to $\dot{\mu}_j$ and $\dot{\Sigma}_{jj}$ than those obtained by the pilot run of the GVS (Fig. 5). Thus, the pseudo-priors of our algorithm seem to have been also closer to the marginal posterior distributions of the coefficients than those of the GVS, which may have provided a better mixing of our algorithm than that of the GVS.

We calculated the mean absolute error of variable selection EP and the cross entropy CE of the logistic models estimated by the SFS, the Lasso and the Bayesian variable selection. Both the mean absolute error and the cross entropy for the Bayesian variable selection based on our algorithm were lower than those for the SFS and the Lasso (Table 5).

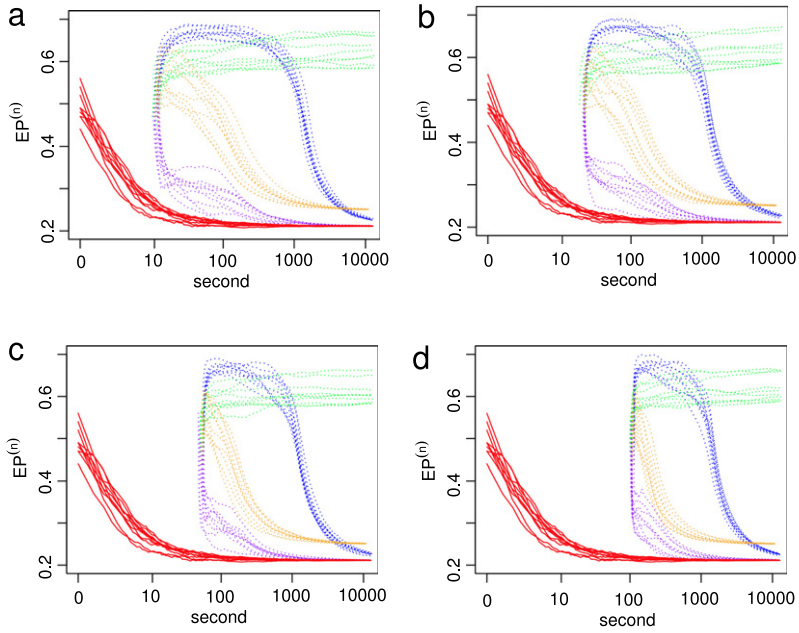

Fig. 2. Trace plots of the $EP^{(n)}$ by the adaptive IMS (red line), the GVS (blue dotted line), the KM method (purple dotted line), the SSVS (marigold dotted line) and the RJMCMC (green dotted line) in 10 runs. The iteration numbers of the pilot runs of the conventional algorithms are (a) 10^3 (b) 2×10^3 (c) 5×10^3 (d) 10^4. The $EP^{(n)}$ by the conventional algorithms are plotted after their pilot runs complete. The computational time is transformed logarithmically. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

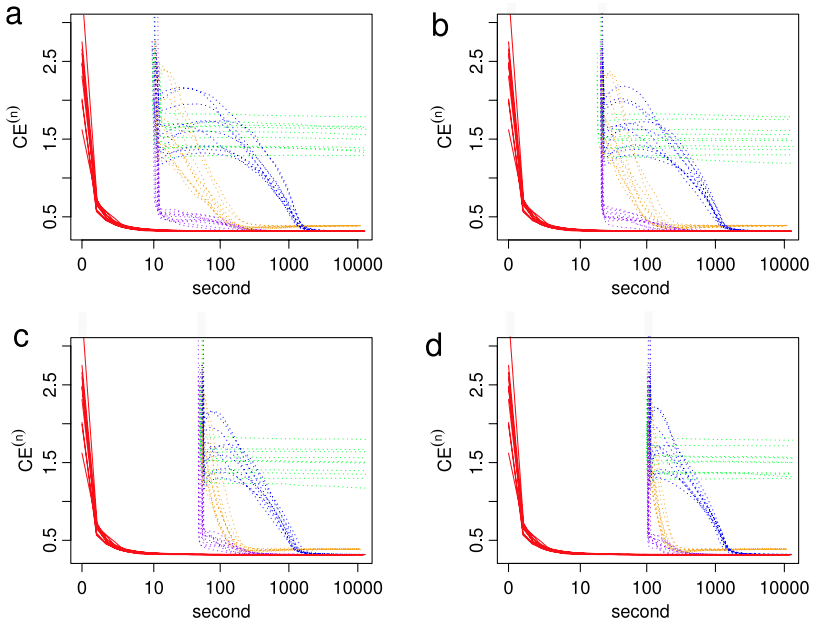

Fig. 3. Trace plots of $CE^{(n)}$ by the adaptive IMS (red line), the GVS (blue dotted line), the KM method (purple dotted line), the SSVS (marigold dotted line) and the RJMCMC (green dotted line) in 10 runs. The iteration numbers of the pilot runs of the conventional algorithms are (a) 10^3 (b) 2×10^3 (c) 5×10^3 (d) 10^4. The computational time is transformed logarithmically. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

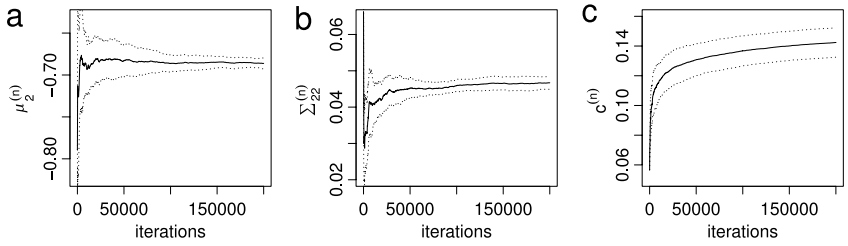

Fig. 4. Trace plot (a) posterior mean $\mu_2^{(n)}$, (b) posterior variance $\Sigma_{22}^{(n)}$, (c) scale parameter $c^{(n)}$. The solid line and the dotted line show mean and mean $\pm$ standard deviation in 10 runs, respectively. The scale $c^{(n)}$ continued to slightly increase from about 2×10^4 to the last iteration, but the slight increase is negligible and does not depress the efficiency of the algorithm.

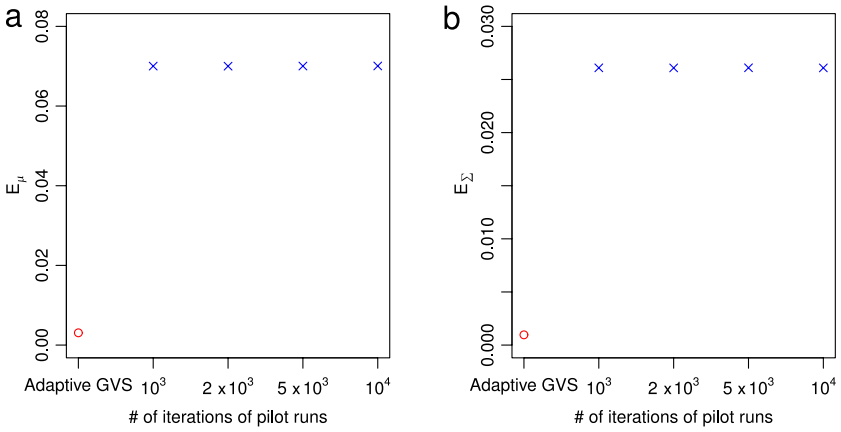

Fig. 5. The averages of E_μ (a) and E_Σ (b) obtained by the 10 runs of the adaptive IMS (red ○) and the pilot runs of the GVS (blue ×) whose iteration numbers are 10^3, 2×10^3, 5×10^3 and 10^4. All of the standard deviations are less than 3×10^{-4} and are negligible. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Averages and standard deviations of the mean acceptance rates in the adaptive IMS, the GVS and the KM method with every pilot run for the posterior of the logistic regression model given the synthetic data. The mean acceptance rates were calculated by Metropolis acceptance results after burn-in period, 2×10^4 iterations. The standard deviations are enclosed by parentheses.

Iteration number of pilot run	10^3	2×10^3	5×10^3	10^4
Adaptive IMS	0.242 (0.92×10^{-3})			
GVS	0.864 (2.3×10^{-3})	0.864 (1.7×10^{-3})	0.864 (1.1×10^{-3})	0.864 (1.4×10^{-3})
KM method	0.864 (2.82×10^{-3})	0.864 (2.78×10^{-3})	0.863 (1.58×10^{-3})	0.864 (1.33×10^{-3})

Table 5
The mean absolute errors and the cross entropies for the Lasso, the SFS and the Bayesian variable selection using the adaptive IMS. The values for the Bayesian variable selection were $EP^{(N)}$ and $CE^{(N)}$ obtained by the adaptive IMS (Figs. 2 and 3). Each value for the Bayesian variable selection is the worst value in the 10 runs of our algorithm.

	Mean absolute error	Cross entropy
Bayesian variable selection	0.212	0.316
Lasso	0.380	0.323
SFS	0.290	0.340

5.2.2. Cardiac arrhythmia data

The cardiac arrhythmia data were measured from patients that have cardiac arrhythmia or not, and were used to classify them (Guvenir, Acar, Demiroz, & Cekin, 1997). The data consist of $p = 257$ covariates and 452 instances. The 245 instances have not cardiac arrhythmia, $y = 0$, and the others have, $y = 1$. The covariates that contain missing values were excluded.

To compare the efficiency of the algorithms, we calculated the estimated inefficiency factor

$$\mathrm{IF}_j = 1 + 2 \sum_{i=1}^{M} \left(1 - \frac{i}{m}\right) \hat{\rho}_j(i),$$

where $\hat{\rho}_j(i)$ is an estimated autocorrelation of the sample $\gamma_j^{(n)}$ after burn-in, and M is a truncation point until which the estimated autocorrelation is significant, and m is the number of samples after burn-in. IF_j is proportional to a variance estimator of the sample mean of $\gamma_j^{(n)}$, which is also the estimate of the posterior probability of inclusion.

The adaptive IMS, the GVS and the KM method were run for 4×10^5 iterations. To match the computational times roughly, the SSVS and the RJMCMC were run for 16×10^5 and $4 \times 258 \times 10^5$ iterations, respectively (Table 6). We drew the samples of the SSVS and the RJMCMC every 40 and 2580 iterations, respectively. All the algorithms were run 10 times. The pilot runs of the conventional algorithms were run for 10^4, 5×10^4 and 10^5 iterations. The initial sample values were independently randomly chosen as $\beta_j^{(0)} \sim U(-1 + \hat{\beta}_j, 1 + \hat{\beta}_j)$ and $\gamma_j^{(0)} \sim Be(0.5)$.

Table 6
Averages and standard deviations of the computational times (second) in the adaptive IMS, the GVS, the KM method, the SSVS and the RJMCMC with every pilot run.

Iteration number of pilot run	10^4	5×10^4	10^5
Adaptive IMS	72×10^3 (0.4×10^3)		
GVS	77×10^3 (2.6×10^3)	82×10^3 (1.5×10^3)	86×10^3 (2.4×10^3)
KM method	71×10^3 (1.4×10^3)	80×10^3 (1.4×10^3)	92×10^3 (0.9×10^3)
SSVS	70×10^3 (0.3×10^3)	75×10^3 (0.3×10^3)	81×10^3 (0.4×10^3)
RJMCMC	63×10^3 (0.8×10^3)	68×10^3 (0.7×10^3)	73×10^3 (0.3×10^3)

Fig. 6 shows IF_j in the adaptive IMS and the GVS, the KM method, the SSVS and the RJMCMC with every pilot run except diverged estimates, whose corresponding $\gamma_j^{(n)}$ took the same value every iteration after the burn-in period. All IF_j of our algorithm were lower than those of the conventional algorithms except the RJMCMC for every pilot run, and almost all IF_j of our algorithm were lower than those of the RJMCMC. This indicates the adaptive IMS is more efficient than the conventional algorithms, even if they conduct their pilot runs for many iterations.

The rates in the adaptive IMS were also close to desirable value $\alpha = 0.234$ (Table 7).

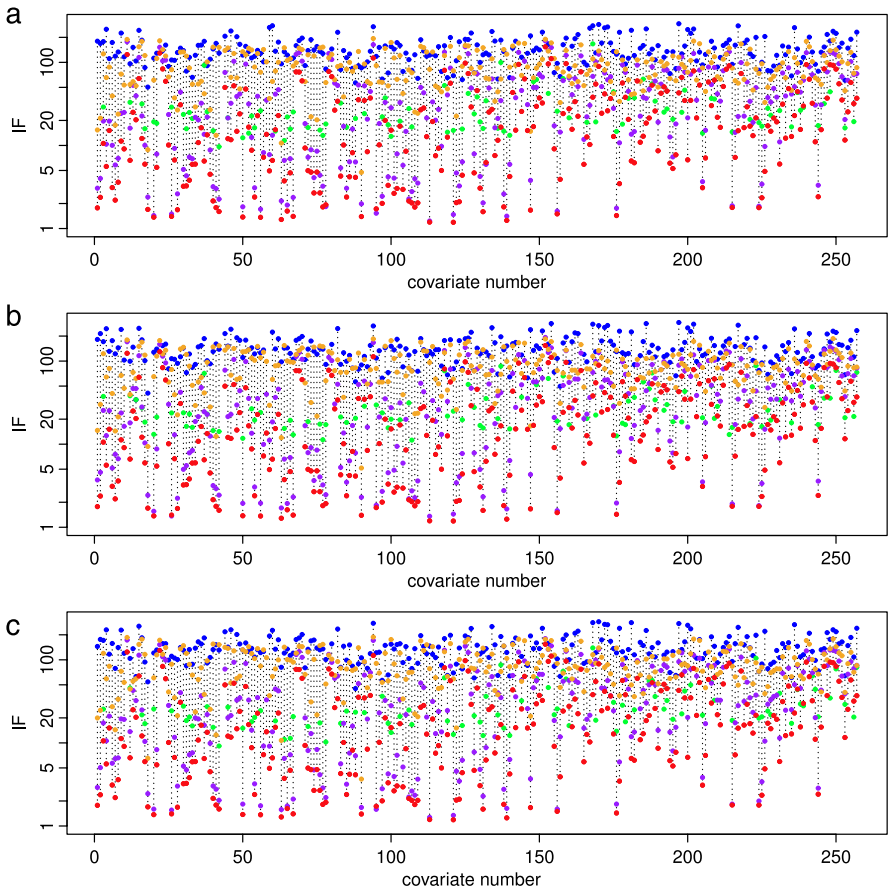

Fig. 6. The mean of IF_j of the adaptive IMS (red points), the GVS (blue points), the KM method (purple points), the SSVS (marigold points) and the RJMCMC (green points) in 10 runs. The iteration number of the pilot runs of the conventional algorithms are (a) 10^4 (b) 5×10^4 (c) 10^5. The IF_j are transformed logarithmically as $\log_{10}(IF_j)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Averages and standard deviations of the mean acceptance rates in the adaptive IMS, the GVS and the KM method with every pilot run for the posterior of the logistic regression model given the cardiac arrhythmia data. The mean acceptance rates were calculated by Metropolis acceptance results after burn-in period, 4×10^4 iterations.

Iteration number of pilot run	10^4	5×10^4	10^5
Adaptive IMS		0.233 (0.2×10^{-3})	
GVS	0.740 (4.7×10^{-3})	0.743 (4.5×10^{-3})	0.739 (2.3×10^{-3})
KM method	0.740 (3.1×10^{-3})	0.741 (2.4×10^{-3})	0.739 (1.7×10^{-3})

Table 8
The 5-fold cross validation errors of the logistic regression probability for the Bayesian variable selection using the adaptive IMS and the RJMCMC, the Lasso and the SFS.

	Bayesian variable selection		Lasso	SFS
	Adaptive IMS	RJMCMC		
Cross validation error	0.329	0.343	0.354	0.373

We calculated 5-fold cross validation errors of the logistic regression probability for the Bayesian variable selection using the adaptive IMS and the RJMCMC, the Lasso and the SFS. The 5-fold cross validation error is $CV = \frac{1}{5}\sum_{k=1}^{5} E_k$, where $E_k = \frac{1}{n_k} \sum_{i=1}^{n_k} |y_i^* - \hat{f}(y = 1|x_i^*)|$, where $\{x_i^*, y_i^*\}_{i=1}^{n_k}$ are validation data, and $\hat{f}$ is estimated logistic model by the training data. The cross validation error for the Bayesian variable selection using the adaptive IMS is lower than those for the Lasso and the SFS (Table 8). The cross validation error obtained by the adaptive IMS is also lower than that by the RJMCMC, because the RJMCMC could not converge even for $258 \times 4 \times 10^5$ iterations.

6. Conclusions

This paper proposed an adaptive algorithm that adapts parameters of a proposal distribution and pseudo-priors during generating samples to overcome the parameter setting problems of the IMS, and proved its convergence theorem. We also showed the proposed algorithm mixes faster than the conventional algorithms through experiments of the Bayesian variable selection of the logistic regression models.

The proposed algorithm enables us to perform the Bayesian variable selection faster than the conventional algorithms. The Bayesian variable selection has following advantages over non-Bayesian variable selection methods such as the Lasso estimator and the SFS. The Bayesian variable selection can estimate not only the important covariates and their coefficients but also the posterior probability that each covariate is included in the model and the predictive distribution. The predictive distribution is a powerful tool widely applied in the classification or the prediction problems. On the other hand, the non-Bayesian variable selection methods can only extract important covariates and estimate their coefficients. Practically, we showed the predictive distribution estimated by the Bayesian variable selection has higher variable selection and generalization ability than the model estimated

by the Lasso and the SFS in the experiments of the normal linear model and the logistic regression model. Furthermore, Chakraborty (2009) showed that the classification model based on the predictive distribution has higher classification performance than the models estimated with L_1 penalty in the classification of gene expression microarray data.

The IMS will be efficient for the Bayesian variable selection of more complex models such as a structural equation model and a non-Gaussian graphical model. We will apply our algorithm to these models and enable more efficient sampling for the Bayesian variable selection of these models.

Acknowledgment

This study was supported by MEXT KAKENHI Grant Number 25120011.

Appendix A. Formulation of the adaptive IMS as the adaptive MCMC for AVMs

The AVMs generate the samples from the joint distribution of the target distribution and the auxiliary distribution which improves the efficiency of sampling from the target distribution. First, we formulate the IMS as the AVMs, and then confirm the adaptive IMS is one of the adaptive MCMC for AVMs.

To formulate the IMS as the AVMs, we regard the IMS generates the samples from the posterior distribution of γ, β and $\theta_j = \gamma_j\beta_j$. Then the posterior distribution is represented as the joint distribution of the following target distribution and auxiliary distribution. The target distribution is the posterior distribution of γ_j and $\theta_j = \gamma_j\beta_j$,

$$f(\theta, \gamma|D) \propto f(D|\theta) \prod_{j=1}^{p} \left(\gamma_j f_j(\theta_j) + (1-\gamma_j)\mathbf{1}_{\{0\}}(\theta_j)\right) f_j(\gamma_j), \tag{A.1}$$

where $\mathbf{1}_{\{\cdot\}}(y)$ is an indicator function, and $f_j(\theta_j)$ is the coefficient prior in Eq. (1). Note that this target distribution has no parameters. The auxiliary distribution is

$$f_\lambda(\beta|\gamma, \theta, D) \propto \prod_{j=1}^{p} \left(\gamma_j \mathbf{1}_{\{\theta_j\}}(\beta) + (1-\gamma_j) f_{\lambda_j}(\beta_j)\right).$$

Since the IMS produces the samples from the joint distribution of the target distribution and the auxiliary distribution, the IMS can be regarded as the AVMs.

The adaptive MCMC for AVMs adapts the parameters of only the proposal distribution and the auxiliary distributions while it runs. The adaptive IMS also adapts the parameters of only those distributions, the means and the variances of the pseudo-priors and the covariance matrix of the proposal distribution, during running. Therefore, the adaptive IMS is included in the adaptive MCMC for AVMs.

Appendix B. Ergodicity

The parameters of the IMS, $\Phi = (\Sigma, \mu, c)$, are contained in the space $\Omega = \Xi \times \mathcal{U} \times \mathcal{C}$. The adaptive IMS is ergodic if and only if

$$\lim_{n\to\infty} \|A^{(n)}((\beta, \gamma, \Phi), (d\theta, d\gamma)) - F(d\theta, d\gamma|D)\| = 0,$$
$$\forall(\beta, \gamma) \in \Theta \times \Gamma,\ \forall \Phi \in \Omega,$$

where $\Gamma = \{0, 1\}^p$, $\|\mu(d\theta, d\gamma) - \nu(d\theta, d\gamma)\| = \sup_{A\in\mathcal{F}_{\Theta\times\Gamma}} |\mu(A) - \nu(A)|$, where $\mathcal{F}_{\Theta\times\Gamma}$ is σ-algebra on $\Theta \times \Gamma$, $F(d\theta, d\gamma|D)$ is the posterior distribution of θ and γ whose density is $f(\theta, \gamma|D)$, Eq. (A.1), and

$$A^{(n)}((\beta, \gamma, \Phi), B_{\theta,\gamma}) = P\left[(\theta^{(n)}, \gamma^{(n)}) \in B_{\theta,\gamma} | \beta^{(0)} = \beta, \gamma^{(0)} = \gamma, \Phi^{(0)} = \Phi\right],$$
$$B_{\theta,\gamma} \in F_{\Theta\times\Gamma}.$$

Appendix C. Proof of Theorem 1

Our proof makes use of Theorem 2 in Araki and Ikeda (2013), which implies that the adaptive IMS is ergodic if it satisfies the following three conditions.

(a) Simultaneously strongly aperiodically geometrical ergodicity
There exists $C \in \mathcal{F}_{\Theta\times\Gamma\times\Theta}$, $V : \Theta \times \Gamma \times \Theta \to [1, \infty)$, $\delta > 0$, $\tau < 1$, and $b < \infty$, such that $\sup_C V < \infty$, $E[V(\beta^{(0)}, \gamma^{(0)}, \theta^{(0)})] < \infty$, and the following conditions hold for all $\Phi \in \Omega$.
(i) (Strongly aperiodic minorization condition)
There exist a probability measure $\nu_\Phi(d\beta, d\gamma, d\theta)$ on C such that
$P_\Phi((\beta, \gamma, \theta), (d\beta', d\gamma', d\theta')) \geq \delta\nu_\Phi(d\beta', d\gamma', d\theta')$, for all $\beta, \gamma, \theta \in C$,
where P_Φ is a transition kernel of the IMS with the parameters Φ.
(ii) (Geometric drift condition)
$(P_\Phi V)(\beta, \gamma, \theta) \leq \tau V(\beta, \gamma, \theta) + b\mathbf{1}_{\{C\}}(\beta, \gamma, \theta)$, for all $\beta, \gamma, \theta \in \Theta \times \Gamma \times \Theta$,
where $(P_\Phi V)(\beta, \gamma, \theta) \equiv \iiint P_\Phi((\beta, \gamma, \theta), (d\beta', d\gamma', d\theta'))V(\beta', \gamma', \theta')$.

(b) Diminishing adaptation

$$\lim_{n\to\infty} \sup_{(\beta,\gamma,\theta)\in\Theta\times\Gamma\times\Theta} \|P_{\Phi^{(n+1)}}\left((\beta, \gamma, \theta), (d\beta', d\gamma', d\theta')\right) - P_{\Phi^{(n)}}\left((\beta, \gamma, \theta), (d\beta', d\gamma', d\theta')\right)\| = 0$$
in probability. (C.1)

First, we prove the condition (a) holds.

By the condition (s1), we have $d_1 \equiv \sup_{(\lambda,\beta,\gamma)\in\Lambda\times S\times\Gamma} f_\lambda(\beta, \gamma|D) < \infty$, $d_2 \equiv \inf_{(\lambda,\beta,\gamma)\in\Lambda\times S\times\Gamma} f_\lambda(\beta, \gamma|D) > 0$ and $\delta_f \equiv \inf_{(\beta,\lambda,\gamma)\in S\times\Lambda\times\Gamma_{\backslash\mathbf{0}}} f_\lambda(\beta_\gamma) > 0$, where $f_\lambda(\beta_\gamma) = \prod_{\beta_j\in\beta_\gamma} f_{\lambda_j}(\beta_j)$ and $\Gamma_{\backslash\mathbf{0}}$ is the set of all elements of Γ except the vector whose all elements are 0. By the compactness of S and the condition (s2), we have $\delta_q \equiv \inf_{(\beta,\beta',\gamma,\Sigma,c)\in S^2\times\Gamma_{\backslash\mathbf{0}}\times\Xi\times\mathcal{C}} q_{(\Sigma_\gamma, c)}(\beta'_\gamma|\beta_\gamma) > 0$. We denote $\delta = \min(\delta_f, \delta_q, 1)$.

For $\beta \in S$, denote $R_\beta = \left\{\beta' \in S, \gamma' \in \Gamma, \theta' = (\beta' \cdot \gamma') \middle| \frac{f_{\gamma'}(\beta'_{\gamma'}|\gamma', D)}{f_{\gamma'}(\beta_{\gamma'}|\gamma', D)} \leq 1\right\}$, where $(\beta' \cdot \gamma') = (\beta'_1\gamma'_1, \ldots, \beta'_p\gamma'_p)$ and $f_\gamma(\beta_\gamma|\gamma, D) = f(D|\beta_\gamma, \gamma)\prod_{\beta_j\in\beta_\gamma} f_j(\beta_j)$.

For $(\beta, \gamma, \theta) \in W \equiv S \times \Gamma \times S$, $\Phi \in \Omega$ and $B \in F_W$, where F_W is σ-algebra on W, we have

$$\begin{aligned}
P_\Phi((\beta, \gamma, \theta), B) &= \iint_{(\beta',\gamma',\theta')\in B} f_\lambda(\gamma'|\gamma, \beta, D) f_\lambda(\beta'_{\backslash\gamma'}) \\
&\quad \times \left\{ q_{(\Sigma_{\gamma'}, c)}(\beta'_{\gamma'}|\beta_{\gamma'}) \min\left(1, \frac{f_{\gamma'}(\beta'_{\gamma'}|\gamma', D)}{f_{\gamma'}(\beta_{\gamma'}|\gamma', D)}\right) \right. \\
&\quad + \mathbf{1}_{\{B\}}(\dot\beta', \gamma', (\gamma' \cdot \dot\beta')) \int_{\{\beta_{\gamma'}|\dot\beta'\in\Theta\}} q_{(\Sigma_{\gamma'}, c)}(\tilde\beta_{\gamma'}|\beta_{\gamma'}) \\
&\quad \left. \times \left(1 - \min\left(1, \frac{f_{\gamma'}(\tilde\beta_{\gamma'}|\gamma', D)}{f_{\gamma'}(\beta_{\gamma'}|\gamma', D)}\right)\right) d\tilde\beta_{\gamma'} \right\} d\beta' d\gamma' \\
&\geq \int_{B\cap R_\beta} f_\lambda(\gamma'|\gamma, \beta, D)\delta^2 \frac{f_{\gamma'}(\beta'_{\gamma'}|\gamma', D)}{f_{\gamma'}(\beta_{\gamma'}|\gamma', D)} d\beta' d\gamma' \\
&\quad + \int_{B\cap R_\beta^c} f_\lambda(\gamma'|\gamma, \beta, D)\delta^2 d\beta' d\gamma' \\
&\geq \left(\frac{d_2}{2d_1}\right)^p \delta^2 \int_{B\cap R_\beta} \frac{f_\lambda(\beta', \gamma'|D)}{f_\lambda(\dot\beta', \gamma'|D)} d\beta' d\gamma'
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{d_2}{2d_1}\right)^p \delta^2 \int_{B\cap R_\beta^c} \frac{f_\lambda(\beta', \gamma'|D)}{f_\lambda(\beta', \gamma'|D)} d\beta' d\gamma' \\
&\geq \frac{d_2^p \delta^2}{2^p d_1^{(p+1)}} \int_{B\cap R_\beta} f_\lambda(\beta', \gamma'|D) d\beta' d\gamma' \\
&\quad + \frac{d_2^p \delta^2}{2^p d_1^{(p+1)}} \int_{B\cap R_\beta} f_\lambda(\beta', \gamma'|D) d\beta' d\gamma' \\
&= \frac{d_2^p \delta^2}{2^p d_1^{(p+1)}} \int_{B} f_\lambda(\beta', \gamma'|D) d\beta' d\gamma',
\end{aligned}
$$

where $f_\lambda(\gamma'|\gamma, \beta, D) = f_\lambda(\gamma'_1|\gamma'_{-1}, \beta, D) f_\lambda(\gamma'_2|\gamma'_{-2}, \beta, D) \cdots f_\lambda(\gamma'_p|\gamma'_{-p}, \beta, D)$, where $\gamma'_{-j} = (\gamma'_1, \ldots, \gamma'_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$, and $\dot{\beta}' = (\dot{\beta}'_1, \ldots, \dot{\beta}'_p)$, where $\dot{\beta}'_j$ is β'_j if $\gamma'_j = 0$, and β_j otherwise. This inequality indicates the condition (a)(i) follows.

Let $0 < \tau < 1$, $V(\beta, \gamma, \theta) = 1$ if $(\beta, \gamma, \theta) \in W$, otherwise $V(\beta, \gamma, \theta) = 1/\tau$, and $b = 1 - \tau$. Then the inequality of the condition (a)(ii) holds for all $\Phi \in \Omega$. That is, the condition (a)(ii) is satisfied. We have also $E[V(\beta^{(0)}, \gamma^{(0)}, \theta^{(0)})] \leq 1/\tau < \infty$.

From the update Eqs. (3) and (4), it follows that $\mu_j^{(n+1)} - \mu_j^{(n)} \to 0$, $\Sigma_{ij}^{(n+1)} - \Sigma_{ij}^{(n)} \to 0$, and $c^{(n+1)} - c^{(n)} \to 0$ as $n \to \infty$. Thus, the condition (b) holds.

The proof is complete.

References

Araki, T., & Ikeda, K. (2013). Adaptive Markov chain Monte Carlo for auxiliary variable method and its application to parallel tempering. *Neural Networks*, *43*, 33–40.

Atchade, Y. (2011). A computational framework for empirical Bayes inference. *Statistics and Computing*, *21*, 463–473.

Chakraborty, S. (2009). Bayesian binary kernel probit model for microarray based cancer classification and gene selection. *Computational Statistics & Data Analysis*, *53*, 4198–4209.

Dellaportas, P., Forster, J. J., & Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, *12*, 27–36.

George, E., & McCullogh, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Guvenir, H., Acar, B., Demiroz, G., & Cekin, A. 1997. A supervised machine learning algorithm for arrhythmia analysis. In *Proceedings of the computers in cardiology conference* (pp. 433–436).

Hastie, T. J., & Pregibon, D. (1992). Generalized linear models. In *Statistical models in S*. (Chapter 6).

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā Series B*, *60*, 65–81.

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

Paroli, R., & Spezia, L. (2007). Bayesian variable selection in Markov mixture models. *Communications in Statistics. Simulation and Computation*, *37*, 25–47.

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

Roberts, G., Gelman, A., & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*, 110–120.

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