Published in final edited form as:
Stat Probab Lett. 2011 February 1; 81(2): 283–291. doi:10.1016/j.spl.2010.10.011.

Bayesian Variable Selection via Particle Stochastic Search

Minghui Shi[a,1] and **David B. Dunson**[a,2]

Minghui Shi: ms193@stat.duke.edu; David B. Dunson: dunson@stat.duke.edu

[a]Department of Statistical Science, Box 90251, Duke University, Durham, NC, 27708, USA

Abstract

We focus on Bayesian variable selection in regression models. One challenge is to search the huge model space adequately, while identifying high posterior probability regions. In the past decades, the main focus has been on the use of Markov chain Monte Carlo (MCMC) algorithms for these purposes. In this article, we propose a new computational approach based on sequential Monte Carlo (SMC), which we refer to as particle stochastic search (PSS). We illustrate PSS through applications to linear regression and probit models.

Keywords

Bayes factor; Marginal inclusion probability; Model averaging; Model uncertainty; Sequential Monte Carlo; Stochastic search variable selection; Subset selection

1. Introduction

Let y_i denote a response variable and $\mathbf{x}_i = (x_{i1},\ldots, x_{ip})'$ denote a $p \times 1$ vector of candidate predictors for subject i, $i = 1,\ldots, N$. Following common notation, let $\gamma_j = 1$ denote that the jth predictor is included in the model with $\gamma_j = 0$ otherwise. Then, $\boldsymbol{\gamma} = (\gamma_1,\ldots, \gamma_p)'$ is a predictor inclusion indicator belonging to a model space Γ, with Γ containing 2^p elements corresponding to all possible subsets of these p candidate predictors. Conditional on $\boldsymbol{\gamma}$, the regression model can be written as

$$(y_i|\mathbf{x}_{\gamma,i}, \gamma, \theta_\gamma) \sim f(\mathbf{x}_{\gamma,i}, \gamma, \theta_\gamma) \quad \text{independently} \tag{1}$$

where $\mathbf{x}_{\boldsymbol{\gamma},i} = \{1, x_{ij}, j : \gamma_j = 1\}$ is the predictor vector, $\theta_{\boldsymbol{\gamma}}$ are the parameters, and $p_\gamma = \sum_{j=1}^{p} \gamma_j$ is the number of predictors in model $\boldsymbol{\gamma}$.

There is a rich literature on methods for sparse point estimation using methods such as Lasso <Tibshirani, 1996>, the relevance vector machine <Tipping, 2001> and the elastic net <Zou and Hastie, 2005>. Although these sparse point estimation approaches often do a good job in simultaneously selecting predictors and estimating the coefficients, they do not allow for uncertainty in variable selection. When p is moderate to large, there is substantial

[1]Minghui Shi is a graduate student in the Department of Statistical Science, Duke University.
[2]David B. Dunson is a professor in the Department of Statistical Science, Duke University.

uncertainty in variable selection, and it is important to allow for this uncertainty in conducting predictions and inferences about the important predictors. To obtain more realistic predictive intervals and potentially lowered mean square predictive error, Bayesian model averaging can be used <Raftery, Madigan, and Hoeting, 1998>. In addition, marginal inclusion probabilities provide a useful measure of the weight of evidence in the data that a particular predictor should be included in the model.

To define a Bayesian approach for variable selection, let $\pi(\boldsymbol{\gamma})$ denote the prior probability of model $\boldsymbol{\gamma}$. Updating this prior with information in the data $\mathbf{y}_{1:N}=\{y_i\}_{i=1}^N$ with $\mathbf{X}_{1:N}=\{\mathbf{x}_i\}_{i=1}^N$, we obtain

$$\pi(\gamma \mid \mathbf{y}_{1:N}, \mathbf{X}_{1:N})=\frac{\pi(\gamma)L(\mathbf{y}_{1:N};\gamma,\ \mathbf{X}_{1:N})}{\sum_{\gamma^*\in\Gamma}\pi(\gamma^*)L(\mathbf{y}_{1:N};\gamma^*,\mathbf{X}_{1:N})} \tag{2}$$

with $L(\mathbf{y}_{1:N};\ \boldsymbol{\gamma},\ \mathbf{X}_{1:N}) = \int L(\mathbf{y}_{1:N};\ \theta_{\boldsymbol{\gamma}},\ \mathbf{X}_{1:N})d\pi(\theta_{\boldsymbol{\gamma}})$ the marginal likelihood under model $\boldsymbol{\gamma}$ and $L(\mathbf{y}_{1:N};\ \theta_{\boldsymbol{\gamma}},\ \mathbf{X}_{1:N})$ the likelihood of $\mathbf{y}_{1:N}$ conditionally on the predictors $\mathbf{X}_{1:N}$ under (1). Expression (2) describes the posterior probabilities for each of the candidate models, with these posterior probabilities providing weights to be used in model averaging or a means by which to conduct Bayesian variable selection.

In particular, if the goal is to select a single "best" model, then there are two approaches that are typically used. First, if one chooses a 0–1 loss function in which a loss of 1 is accrued if an incorrect model is selected and it is assumed that the true model is one of those in the list Γ, then the model with lowest Bayes risk corresponds to the highest posterior probability model. Examining expression (2), the posterior probability of model $\boldsymbol{\gamma}$ is proportional to the prior probability multiplied by the marginal likelihood under that model. Due to the intrinsic Bayesian penalty for model dimension <Jefferys and Berger, 1992>, the marginal likelihood will tend to favor a parsimonious model. However, there are two major problems that arise in selecting the highest posterior probability model when 2^p is large. First, the number of models that need to be visited in calculating the denominator in (2) rapidly becomes prohibitively large as p increases, and hence it becomes difficult to accurately estimate $\pi(\boldsymbol{\gamma}|\ \mathbf{y}_{1:N},\ \mathbf{X}_{1:N})$. Second, even if an exact estimate could be obtained, no one model will dominate in large model spaces, and it tends to be the case that many models have similar posterior probabilities to the best model.

To address these problems, it has become common to instead select predictors based on thresholding of the marginal inclusion probabilities (MIPs), defined as

$$\zeta_j=P(\gamma_j=1 \mid \mathbf{y}_{1:N}, \mathbf{X}_{1:N})=\sum_{\gamma:\gamma_j=1}\pi(\gamma \mid \mathbf{y}_{1:N}, \mathbf{X}_{1:N}) \tag{3}$$

for the jth predictor, $j=1,\ldots, p$. The MIPs provide a weight of evidence that a given predictor should be included adjusting for uncertainty in the other predictors in the model, and hence provide a useful basis for inferences. Barbieri and Berger <2004> showed that the optimal predictive model under squared error loss often corresponds to the median probability model, which includes all predictors having MIPs above 0.5. Because it is often not feasible to visit more than a small fraction of the models in Γ in estimating the MIPs, it is important to develop algorithms that efficiently find regions of Γ containing high posterior probability models, with such models also tending to have high marginal likelihoods unless the prior is overly informative.

George and McCulloch <1993> proposed a stochastic search variable selection (SSVS) algorithm for normal linear regression using Gibbs sampling to search Γ for high posterior probability models. Their approach relies on a mixture of a low and high variance normal prior centered at zero for each of the regression coefficients, with the low variance component corresponding to a predictor being effectively excluded due to the coefficient being close to zero. However, in many applications, this approach is subject to very slow mixing of the Gibbs sampler and hence poor computational efficiency <George and McCulloch, 1997>. As reviewed in George and McCulloch <1997>, Geweke <1996>, Carlin and Chib <1995> and Green <1995> propose alternative methods to improve the performance of SSVS. As noted in Liu *et al.* <1994>, an effective strategy for improving efficiency of MCMC algorithms is marginalization. The most efficient of the available SSVS algorithms (to our knowledge) relies on marginalizing out the regression coefficients in updating the variable inclusion indicators <George and McCulloch, 1997>. In particular, this algorithm iteratively samples the variable inclusion indicator for the *j*th predictor, γ_j, from its Bernoulli full conditional posterior distribution given the other predictors in the model, $\gamma_{(-j)} = \{\gamma_l : l \neq j, l = 1,\ldots, p\}$, for $j = 1,\ldots p$.

In this article, we propose a sequential Monte Carlo (SMC) approach for obtaining a sampling-based approximation to the posterior distribution of $\boldsymbol{\gamma}$, providing an alternative to SSVS and other MCMC-based methods. Although SMC is commonly used for dynamic models, the application to static models was initially proposed by Chopin <2002>, with Del Moral *et al.* <2006> providing a general methodology. However, there has been limited work on the use of SMC for model selection. Chopin <2007> used SMC for model choice in hidden Markov models. Toni *et al.* <2009> proposed an approximate Bayesian computation method for model selection in dynamical systems using SMC. Zhang *et al.* <2007> proposed an SMC-type sequential optimization approach for variable selection, though their approach does not accommodate uncertainty in the selection process.

Our proposed particle stochastic search (PSS) algorithm relies on introducing a sequence of particle approximations to the partial posterior distributions $\{\pi(\gamma \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})\}_{n=1}^{N}$, with the particles sequentially updated through rejuvenating and reweighing operations as subjects are added to the data set. By adding data sequentially, we initially allow faster exploration of the model space, as the partial posteriors will be effectively annealed relative to the eventual target. In addition, the algorithm can take advantage of distributed computing on a cluster for more rapid computation. In the sequel, we provide details on the PSS approach and compare it to MCMC algorithms in linear and probit regression.

2. Particle Stochastic Search

Due to the dimensionality problem mentioned in Section 1, we focus primarily on obtaining accurate estimates of the MIPs, though the proposed algorithm can also be used to identify high posterior probability models, as we illustrate in Section 3.

2.1. Sequential Monte Carlo for Variable Selection

Sequential Monte Carlo (SMC) relies on a discrete approximation to the posterior distribution

$$\pi(\gamma \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n}) \approx g(\gamma; \{\gamma_m^n, w_m^n\}_{m=1}^{M}) = \sum_{m=1}^{M} w_m^n \delta_{\gamma_m^n}(\gamma), \tag{4}$$

where $\{\gamma_m^n\}_{m=1}^M$ is a collection of particles, $\delta_{\boldsymbol{\gamma}}$ denotes a degenerate distribution with all its mass at $\boldsymbol{\gamma}$, and w_m^n is the probability on particle γ_m^n.

Based on the theory of importance sampling <e.g Liu, 2001, Ch.2>, given a particle approximation $g(\gamma;\{\gamma_m^{n-1}, w_m^{n-1}\}_{m=1}^M)$ to the partial posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:n-1}, \mathbf{X}_{1:n-1})$, one can obtain a particle approximation $g(\gamma;\{\gamma_m^n, w_m^n\}_{m=1}^M)$ to the partial posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$ by propagating the particles $\gamma_m^n=\gamma_m^{n-1}$ and using modified weights

$$w_m^n=\frac{\pi(\gamma_m^n \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})}{\pi(\gamma_m^{n-1} \mid \mathbf{y}_{1:n-1}, \mathbf{X}_{1:n-1})} w_m^{n-1}, m=1, \ldots, M. \tag{5}$$

One can start by drawing $\gamma_m^0 \sim \pi(\gamma)$, $m=1, \ldots, M$, choosing equal weights $\{w_m^0=1/M\}_{m=1}^M$ to obtain the initial approximation, and then apply (5) recursively to obtain a particle approximation (4) for the posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:N}, \mathbf{x}_{1:N})$. However, this sequential weight-updating step has the problem that after several iterations, fewer and fewer particles maintain significant weights. To address this degeneracy problem, a common strategy is to remove particles with very low weights by weighted resampling from $\{\gamma_m^{n-1}\}_{m=1}^M$. Unfortunately, resampling does not introduce new particles, so this approach leads to few particles having very high weight.

Let $K(\boldsymbol{\gamma}^\star \mid \boldsymbol{\gamma})$ denote a transition kernel with invariant probability distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$,

$$\pi(\gamma^\star \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})=\int K(\gamma^\star \mid \gamma)\pi(\gamma|\mathbf{y}_{1:n}, \mathbf{X}_{1:n})d\gamma. \tag{6}$$

Given an initial particle approximation $g(\gamma;\{\gamma_m, w_m\}_{m=1}^M)$, one can use the modified approximation $\sum_{m=1}^M w_m K(\gamma_m^\star \mid \gamma_m)$. To draw samples from this approximated distribution, one can first draw a set of indicators $\{I_m\}_{m=1}^M$ indicating which $\boldsymbol{\gamma}_m$ should be used for the generation of $\gamma_m^\star$, and then sample $\gamma_m^\star$ from $K(\gamma_m^\star \mid \gamma_{I_m})$. The first stage is effectively resampling and the second step allows the generation of fresh particles <Pitt and Shephard, 1999>, <Carvalho *et al.* 2010>.

We consider the following choices of the transition kernel $K(\boldsymbol{\gamma}^\star \mid \boldsymbol{\gamma})$:

1. Metropolis Hasting kernel:
   a. Generate a candidate $\boldsymbol{\gamma}^\star$ from probability distribution $q(\boldsymbol{\gamma}^\star; \boldsymbol{\gamma})$.
   b. Accept the candidate $\boldsymbol{\gamma}^\star$ with probability

$$\alpha(\gamma, \gamma^\star)=\min\left\{1, \frac{q(\gamma;\gamma^\star)}{q(\gamma^\star;\gamma)} \frac{L(\mathbf{y}_{1:n};\gamma^\star, \mathbf{X}_{1:n})\pi(\gamma^\star)}{L(\mathbf{y}_{1:n};\gamma, \mathbf{X}_{1:n})\pi(\gamma)}\right\}. \tag{7}$$

2. Gibbs sampling transition kernel: Let $\gamma_{(j)} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$, and $\tau=(\tau_1, \ldots, \tau_p)$ denote some permutation of $\{1, 2, \ldots, p\}$. Then for $j = \tau_1, \tau_2, \ldots, \tau_p$:

$$(\gamma_j \mid \gamma_{(j)}) \sim \text{Bernoulli}(\widehat{p}_j) \tag{8}$$

with

$$\widehat{p}_j = \frac{\pi(\gamma_{(j)}, \gamma_j = 1 \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})}{\pi(\gamma_{(j)}, \gamma_j = 0 \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n}) + \pi(\gamma_{(j)}, \gamma_j = 1 \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})}.$$

In order to introduce more fresh particles which have less dependence with the previous particles, we can use strategies commonly used for improving the convergence of MCMC. For example, one can use a blocked Gibbs sampling transition kernel. In the following algorithm, the Metropolis Hastings kernel is applied within a particle iteratively *p* times.

In choosing between transition kernels, a useful measure of the efficiency is the effective sample size (ESS), defined as

$$\text{ESS}(N) = \frac{N}{1 + \text{var}(w)}, \tag{9}$$

with var(*w*) the variance of the importance weights with respect to the proposal distribution. The ESS(N) provides an estimate of the number of independent samples from the target probability measure, which would provide the same estimation precision as the particle approximation. It is common to only resample when ESS(N) becomes low.

We propose two alternative PSS algorithms below.

Algorithm 1.

- **i.** *Initialization: Start with sampling the particles* $\{\gamma_m^0\}_{m=1}^M$ *from the prior distribution* $\pi(\boldsymbol{\gamma})$.
- **ii.** *For* $n = 1, 2, \ldots, N$, *add the* n^{th} *observation* $(y_n, \mathbf{x}_n)$ *and cycle through*
  - **a.** *Reweighting: update the weights of the particles:*

$$w_m^n \propto L(y_n; \gamma_m^n, \mathbf{x}_n) \cdot w_m^{n-1} \tag{10}$$

    and set $\gamma_m^n = \gamma_m^{n-1}$ *for* $m = 1, \ldots, M$.
  - **b.** *Calculate the ESS(M)* (9) *based on the updated weights. Once ESS(M)* < *M*=2:
    - **b1)** *(Resample) Resample* $\{\gamma_m^n\}_{m=1}^M$ *with replacement using weights* $\{w_m^n\}_{m=1}^M$ *using an efficient sampling strategy. Reset the weights* $\{w_m^n\}_{m=1}^M$ *to* $\{w_m^n = 1/M\}_{m=1}^M$.
    - **b2)** *(Rejuvenation) For any m, replace* γ_m^n *with a sample from* $K_n(\gamma, \gamma_m^n)$ *where* $K_n(\cdot \mid \boldsymbol{\gamma})$ *defines a transition kernel with invariant probability distribution* $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$.

Del Moral *et al.* <2006> proposed alternatives to sequential adding of observations. For simplicity and to facilitate extensions, we do not consider such approaches here.

2.2. Generalization to Latent Variable Models

For the normal linear regression model, the marginal likelihood is available in closed form when $\pi(\theta\boldsymbol{\gamma})$ is chosen as a multivariate normal-gamma prior. However, for generalized linear models, the marginal likelihood is typically analytically intractable. Albert and Chib <1993> and Holmes and Held <2006> demonstrated auxiliary variable approaches for binary regression models. In this section, we describe the modification of the auxiliary variable approach to our PSS algorithm.

To begin, consider a probit regression model

$$y_i \sim \text{Bernoulli}(\Phi(\eta_i)), \eta_i = \mathbf{x}_i\beta, \beta \sim \pi(\beta), \tag{11}$$

with $\Phi(\cdot)$ the cumulative distribution function of a standard normal random variable. A well known augmented formulation for Model (11) is

$$y_i = \begin{cases} 1 & z_i > 0 \\ 0 & \text{otherwise} \end{cases}, z_i = \mathbf{x}_i'\beta + \varepsilon_i, \varepsilon_i \sim N(0,1),\ \beta \sim \pi(\beta). \tag{12}$$

The advantage of (12) is that, for Gaussian $\pi(\beta)$, we can obtain the marginal likelihood conditionally on the latent variables $\mathbf{z}$ but marginalizing out β. Thus, we can extend our PSS algorithm to probit regression models by including the model index $\boldsymbol{\gamma}$ and the latent variables $\mathbf{z}$ within the particles.

Theorem 1. (*Liu* <2001>) *Let $\pi_0(x, y)$ and $\pi_1(x, y)$ be two probability densities, where the support of π_0 is a subset of the support of π_1. Then,*

$$var_{\pi_1}\left\{\frac{\pi_0(x,y)}{\pi_1(x,y)}\right\} \geq var_{\pi_0}\left\{\frac{\pi_0(x)}{\pi_1(x)}\right\}, \tag{13}$$

where $\pi_1(x) = \int \pi_1(x, y)dy$ and $\pi_0(x) = \int \pi_0(x, y)dy$ are marginal densities.

Based on Theorem 1, we should obtain better performance of the PSS method by avoid putting in the regression parameters specific to each model within the particles and instead marginalizing out these parameters. Marginalization is a common technique for reducing autocorrelation in MCMC algorithms; for example, refer to Holmes and Held <2006> in the setting of SSVS using data augmentation in binary response models.

Let $K_n(\mathbf{z}^\star_{1:n}, \gamma^\star | \mathbf{z}_{1:n}, \gamma)$ denote a transition kernel with invariant distribution $\pi(\mathbf{z}_{1:n}, \boldsymbol{\gamma} \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$ which can be factorized as

$$K_n(\mathbf{z}^\star_{1:n}, \gamma^\star \mid \mathbf{z}_{1:n}, \gamma) = K_n^\gamma(\gamma^\star, \mid \mathbf{z}^\star_{1:n}, \gamma) K_n^z(\mathbf{z}^\star_{1:n} \mid \mathbf{z}_{1:n}, \gamma). \tag{14}$$

We consider the following choice of $K_n(\gamma^\star, \mathbf{z}^\star_{1:n} \mid \mathbf{z}_{1:n}, \gamma)$

1. Gibbs sampling kernel:

$$K_n^z(\mathbf{z}_{1:n}^\star \mid \mathbf{z}_{1:n}, \gamma) = p(z_1^\star \mid \mathbf{z}_{2:n}, \gamma, \mathbf{y}_{1:n}, \mathbf{X}_{1:n}) \prod_{i=2}^{n} p(z_i^\star \mid \mathbf{z}_{1:(i-1)}^\star, \mathbf{z}_{(i+1):n}, \gamma, \mathbf{y}_{1:n}, \mathbf{X}_{1:n}). \tag{15}$$

For probit regression models with a Gaussian prior, we can directly sample from $p(z_i^\star \mid \mathbf{z}_{1:(i-1)}^\star, \mathbf{z}_{(i+1):n}, \gamma, \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$ which is a truncated normal distribution.

2. Gibbs sampling or Metropolis Hasting kernels for $K_n^\gamma(\gamma^\star \mid \mathbf{z}_{1:n}^\star, \gamma)$ as in Section 2.1.

Algorithm 2

i. *Initialization: Start with sampling the particles* $\{\gamma_m^0\}_{m=1}^M$ *from the prior distribution* $\pi(\gamma)$.

ii. *For* $n=1,\ldots, N$, *add the observation* $(y_n, \mathbf{x}_n)$ *and cycle through:*

   a. *Reweighting: update the weights of the particles:*

$$w_m^n \propto L(y_n; \mathbf{z}_{1:n-1,m}^{n-1}, \gamma_m^{n-1}, \mathbf{x}_n) \cdot w_m^{n-1} \tag{16}$$

   and set $(\mathbf{z}_{1:n-1,m}^n, \gamma_m^n) = (\mathbf{z}_{1:n-1,m}^{n-1}, \gamma_m^{n-1})$ *for* $m=1,\ldots, M$.

   b. *Propagating: Sample the next latent variable* z_n *for each particle* m:

$$(z_{n,m}^n \mid z_{1:n-1,m}^n, \gamma_m^n, y_n, \mathbf{x}_n) \sim p(z_n \mid \mathbf{z}_{1:n-1,m}^n, \gamma_m^n, y_n, \mathbf{x}_n) \tag{17}$$

   with the particle system updated to $\{\mathbf{z}_{1:n,m}^n, \gamma_m^n, w_m^n\}_{m=1}^M$.

   c. *Calculate the ESS(M)* (9) *based on the updated weights. If ESS(M)* $< M=2$:

      c1) *(Resample) Resample* $\{\mathbf{z}_{1:n-1,m}^n, \gamma_m^n, w_m^n\}_{m=1}^M$ *with replacement using weights* $\{w_m^n\}$ *based on an efficient sampling strategy. Reset the weights* $\{w_m^n\}_{m=1}^M$ *to* $\{w_m^n = 1/M\}_{m=1}^M$.

      c2) *(Rejuvenation) For any* m, *replace* $(\mathbf{z}_{1:n,m}^n, \gamma_m^n)$ *with a sample from a transition kernel with the invariant distribution* $\pi(\mathbf{z}_{1:n}, \gamma \mid \mathbf{y}_{1:n}, \mathbf{X}_{1:n})$.

Compared with MCMC, PSS has the advantage of avoiding mixing problems, such as a tendency to remain for long intervals within a local region of the model space Γ. However, the tradeoff in SMC algorithms such as PSS is the risk of degeneracy and the potential need to use enormous numbers of particles to obtain an accurate approximation. It is straightforward to extend PSS beyond linear regression and probit models to other models in which marginal likelihoods are available analytically after augmentation. For example, the nonparametric mixture regression models of Chung and Dunson <2009> fall in this class. PSS can be implemented either in serial or in parallel, though a primary advantage of PSS is the ability to accommodate high-dimensional cases through the use of parallel computing.

2.3. Prior Specification and Extensions

The PSS algorithms described above assume that the marginal likelihood $L(\mathbf{y}_{1:N}; \gamma, \mathbf{X}_{1:N})$ can be obtained in closed form, which places some constraints on the priors that can be

considered. For example, in normal linear regression, we have assumed that a multivariate normal-gamma joint prior is placed on the regression coefficients and residual precision. This is a standard choice in the literature. SSVS algorithms that rely on marginalizing out the model parameters also require a closed form marginal likelihood, so have similar restrictions on the prior. For both PSS and SSVS, the class of priors and models that can be considered can be expanded by using approximations to the marginal likelihood, such as Laplace.

There are some disadvantages of the multivariate normal-gamma prior, such as lightness of the tails leading to lack of robustness. A number of alternative priors have been proposed, which place hyper-priors on one or more parameters in the multivariate normal-gamma prior. One example is the mixture of g-priors considered in Liang *et al.* <2008>. In MCMC-based SSVS algorithms, it is straightforward to include hyper-priors on parameters that are common to the different models, and then update these parameters in separate steps from the model index updating steps. In PSS, we can similarly allow richer classes of priors by including the hyper-parameters ψ common to the different models directly in the particles along with the model index $\boldsymbol{\gamma}$. The algorithm would remain essentially the same as described above, but in the rejuvenation step we would need to apply an invariant transition kernel for the joint posterior of $(\boldsymbol{\gamma}, \psi)$. For example, we could use a Gibbs transition kernel.

2.4. Bayesian Inference from the Particles

As described in Section 1, there are a variety of approaches available for selecting predictors based on posterior model probabilities, with our emphasis here being on the median probability model that selects those predictors having marginal inclusion probabilities (MIPs) greater than 0.5. However, in many applications it is not necessary to formally select predictors and may be more useful to present a ranked list of the predictors having the highest MIPs along with their MIPs. As the MIPs provide a weight of evidence that a variable should be included as a predictor, such a summary provides more information than simply a list of selected predictors.

After obtaining a particle approximation $g(\gamma;\{\gamma_m^N, w_m^N=1\}_{m=1}^M)$ to the complete posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}_{1:N}, \mathbf{X}_{1:N})$ over the model space using PSS, the MIP for the jth predictor can be estimated as

$$\widehat{\zeta}_j=\frac{1}{M}\sum_{m=1}^{M}1(\gamma_{m,j}^N=1). \tag{18}$$

After selecting a model based on thresholding of the estimated MIPs, the posterior distribution of the coefficients and residual variance in the selected model can be obtained easily.

3. Examples

3.1. Normal Linear model

We illustrate PSS and compare results to SSVS using simulated examples with the first example taken from George and McCulloch <1997>. To calculate the marginal likelihood $L(\mathbf{y}_{1:N};\boldsymbol{\gamma}, \mathbf{X}_{1:N})$ in both methods, we use a simple multivariate normal-gamma prior distribution for $\theta_{\boldsymbol{\gamma}}$, which includes both the regression coefficients $\beta_{\boldsymbol{\gamma}}$ and the residual precision σ^{-2} in the linear regression case,

$$(\beta_\gamma \mid \sigma^2, \gamma) \sim N(0, \sigma^2 I_\gamma), (\sigma^2 \mid \gamma) \sim \text{IG}(p_\gamma/2, p_\gamma \lambda_\gamma/2),$$

with $\lambda_\gamma = s^2_{LS}$ equal to the classical least square estimate of σ^2 based on the full model as an empirical Bayes approach to set the scale <George and McCulloch, 1997>. In addition, we assumed that the elements of $\boldsymbol{\gamma}$ are iid Bernoulli(0:5) in order to assign equal prior probability to inclusion or exclusion of each predictor.

Algorithms that efficiently discover models with high log-marginal likelihoods will also tend to find models with high posterior probabilities, because the posterior probability is proportional to the prior probability times the marginal likelihood. Hence, we record the log marginal likelihoods of the models visited by PSS and SSVS as one measure of performance, while also estimating MIPs for each of the predictors and the median probability model. Ideally, the MIPs would be close to one for predictors that should be included and close to zero for predictors that should be excluded. However, when important predictors are highly correlated, the MIPs for these predictors will tend to be substantially less than one and may even be less than 0.5. Bayesian variable selection automatically attempts to find a parsimonious model that has good predictive performance, and from this perspective it is often optimal to select one of a correlated set of predictors. The outcomes from all the simulations below are standardized to have mean 0 and unit variance. All of the algorithms are coded in C++, with the PSS algorithms implemented using parallel computation.

Example 1.

Generate $Z_1, Z_2, \ldots, Z_{15}, Z$ from $N_{100}(\mathbf{0}, I)$, and set the covariate X_i to satisfy $X_i = Z_i + 2Z$ for i=1,3,5,8,9,10,12,13,14,15 with $X_2 = X_1 + 0.15Z_1$, $X_4 = X_3 + 0.15Z_4$, $X_6 = X_5 + 0.15Z_6$, $X_7 = X_8 + X_9 - X_{10} + 0.15Z_7$ and $X_{11} = X_{14} + X_{15} - X_{12} - X_{13} + 0.15Z_{11}$. The regression coefficients are $\beta = (1.5, 0, 1.5, 0, 1.5, 0, 1.5, 1.5, 0, 0, 1.5, 1.5, 1.5, 0, 0)'$. The final observation variables are drawn from $y_i \sim N(\mathbf{x}_i'\beta, \sigma^2)$ with σ^2=2.5. Under this construction, there is a strong multicollinearity among the predictors and the correlations between X_i and X_{i+1} are as high as 0.998.

We start with this simple example in order to test the performance of PSS in a case in which the true posterior model probabilities and marginal inclusion probabilities (MIPs) can be calculated precisely. As there are $2^{15} = 32,768$ models in Γ in this case, it is feasible to calculate the marginal likelihood for every model in the list. For a short run, we set the initial number of particles for PSS to be M=1000. However, our hope is that we can still obtain reasonably accurate estimates of the MIPs and identify many of the top posterior probability models based on a modest number of particles. SSVS also typically relies on many fewer samples than there are models in Γ. Matching implementation time approximately, we ran SSVS for 5000 iterations.

We obtained 50 simulation replicates in order to judge performance across many data sets. For each simulation, both PSS and SSVS found the true highest posterior probability model. Since the true MIPs can be obtained in this case, we can calculate root mean square errors (RMSE) of the estimated MIPs and other summaries of performance. Let $\hat{\boldsymbol{\gamma}}_E(\alpha) = \{1(\zeta_j > \alpha), j = 1, \ldots, p\}$ denote the model selected by including predictors having exact MIPs larger than a threshold of α. To assess the relative performance of PSS and SSVS at efficiently approximating exact Bayesian variable selection, we use the following two summaries

$$\mathrm{Ip}(\alpha)=\frac{\sum_{j=1}^{p}1(\zeta_j>\alpha)1(\widehat{\zeta}_j>\alpha)}{\sum_{j=1}^{p}1(\zeta_j>\alpha)},\quad \mathrm{Ep}(\alpha)=\frac{\sum_{j=1}^{p}1(\zeta_j<\alpha)1(\widehat{\zeta}_j<\alpha)}{\sum_{j=1}^{p}1(\zeta_j<\alpha)}$$

with ζ_j and $\hat{\zeta}_j$ the true and estimated MIPs for predictor j, respectively. Here, Ip(α) denotes the proportion of predictors in model $\hat{\boldsymbol{\gamma}}_E(\alpha)$ that are appropriately included in the model selected using the estimated MIPs, while Ep(α) denotes the proportion of predictors not in model $\hat{\boldsymbol{\gamma}}_E(\alpha)$ that are appropriately excluded. Table 1 shows summaries of the RMSE of the estimated MIPs, the means of the Ip and Ep with the standard deviations in the parentheses for both PSS and SSVS. Both PSS and SSVS have excellent performance in terms of accurately approximating Bayesian variable selection based on thresholding of the exact MIPs.

By adding 85 predictors $X_{16:100}$ with zero coefficients, we extended p from 15 to 100 and reapplied PSS and SSVS. In this case the number of models in Γ is too large to calculate the marginal likelihood for all models, so the highest posterior probability model and MIPs cannot be calculated exactly. Hence, we instead compare the relative performance of PSS and SSVS in identifying high log-marginal likelihood models, in estimating MIPs that are high for predictors that should be in the model and in estimating a median probability model that is close to the true model. Table 2 gives the median, 75th percentile, 95th percentile and maximum for the log-marginal likelihoods found in those methods. In this high dimensional case, PSS with 10,000 particles finds slightly higher posterior probability regions than 20,000 SSVS iterations. Table 3 shows the indexes of the predictors in the estimated models based on different thresholdings. The model selected is sensitive to the choice of the thresholding α, with α = 0.5 often an optimal choice in terms of predictive performance <Barbieri and Berger, 2004>.

3.2. Probit Regression Model

We also apply PSS to the following probit regression model with details listed in the Appendix. The prior distribution we used for $\beta_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}$ is $N(\mathbf{b}_{\boldsymbol{\gamma}}, \mathbf{v}_{\boldsymbol{\gamma}})$ with $\mathbf{b}_{\boldsymbol{\gamma}} = 0$ and $\mathbf{v}_{\boldsymbol{\gamma}}=I_{p_{\boldsymbol{\gamma}}\times p_{\boldsymbol{\gamma}}}$ in examples.

Example 2:

Choose the covariate matrix $X^{(p)}=(\mathbf{1}, X)$ with X the same as the covariate matrix in the normal linear regression example with 100 predictors. The response variables are drawn from model (11) with $\eta_i=\mathbf{x}^{(\mathbf{p})'}{}_i\beta^{(\mathbf{p})}$, $z_i \sim N(\eta_i, 1)$, $y_i=1(z_i \geq 0)$ and $\beta^{(\mathbf{p})} = (1.5, \beta)$ with β also being taken from p=100 normal linear regression example.

As the marginal likelihood for the probit model is not available analytically, we instead use the complete data marginal likelihood here, which is available for the simulation as we have generated **z**. In this example, we compare our PSS algorithm with MCMC. As is illustrated in Table 4, PSS with 10,000 particles found slightly higher posterior regions than 20,000 MCMC iterations. The models selected based on different thresholds α on the MIPs are listed in Table 5. If the model selected is (1, 2, 3), it corresponds to $\eta_i=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ in (11).

4. Conclusion

This article has proposed an SMC algorithm for Bayesian variable selection. Our goal in using SMC was to obtain an alternative to MCMC-based SSVS, which may have advantages

in certain cases. First, the proposed PSS algorithm has an automatic annealing feature that results from the sequential addition of subjects. This annealing leads to more rapid exploration of the model space initially and then a more concentrated search as subjects are added. Although annealing is also commonly used within MCMC algorithms to limit problems with stickiness when the posterior is multimodal, the performance of such algorithms tends to be quite sensitive to difficult to choose tuning parameters, such as temperature ladders. PSS incorporates an implicit temperature sequence through making the target more concentrated as subjects are added, so avoids the need to choose tuning parameters.

A second beneficial feature of PSS is that the approach can take advantage of parallel computing environments to simultaneously explore many regions of the model space starting with widely-dispersed particles sampled from the prior. This tends to limit the chance of getting stuck for long intervals in a local region of the model space, and makes it more likely to discover promising regions. Unlike simply implementing SSVS in parallel, PSS automatically communicates across the particles and will discard particles in unpromising low probability regions. Our simulation in the $p = 100$ case provided some initial evidence that PSS has better performance in finding the top models in large predictor spaces, though more extensive simulations and theoretical studies are needed.

This article is meant as an initial description of a promising new class of algorithms for a very challenging and important problem. Certainly, the challenges of attempting posterior computation in a model space with 2^p elements when p is large should not be underestimated. PSS is by no means a perfect alternative to SSVS in that accurate approximation of the posterior of the model index $\boldsymbol{\gamma}$ when p is large would seem to necessitate using an enormous number of particles, which may not be computationally feasible. However, MCMC faces a similar problem in requiring an infeasible number of samples. Hence, it is important to keep in mind that these algorithms are designed to search for good models and not to accurately approximate the posterior for large p. Our hope is that the current PSS algorithm will provide a competitive alternative to SSVS that does better in certain applications, while stimulating additional work in this area. In particular, we suspect that more efficient transition kernels can potentially be chosen to improve performance of PSS.

Appendix

PSS implementation details for the probit model under the prior $\beta \sim N(\mathbf{b}, \mathbf{v})$. Define

$$B_{\gamma,n}=V_{\gamma,n}(\mathbf{v}_{\gamma}^{-1}\mathbf{b}_{\gamma}+\mathbf{X}_{1:n,\gamma}^{\prime}\mathbf{z}_{1:n}),\ V_{\gamma,n}=(\mathbf{v}^{-1}+\mathbf{X}_{1:n,\gamma}^{\prime}\mathbf{X}_{1:n,\gamma})^{-1},$$

1. Reweighting: After marginalizing out the latent variable for the new observations, the weight (16) satisfies

$$w_m^n \propto [1(y_n=1)\{1-\Phi(0\,|\,\widehat{\mu}_n,\widehat{\sigma}^2)\}+1(y_n=0)\Phi(0\,|\,\widehat{\mu}_n,\widehat{\sigma}^2)]w_m^{n-1}$$

with

$$\begin{aligned}\widehat{\mu}_m^n &= \mathbf{x}_{\gamma_m^{n-1},n}^{\prime}V_{\gamma_m^{n-1},n}(\mathbf{v}_{\gamma_m^{n-1}}^{-1}\mathbf{b}_{\gamma_m^{n-1}}+\mathbf{X}_{1:n,\gamma_m^{n-1}}^{\prime}\mathbf{z}_{1:n,m}^{n-1}),\\ (\widehat{\sigma}^2)_m^n &= \mathbf{x}_{\gamma_m^{n-1},n}^{\prime}V_{\gamma_m^{n-1}}\mathbf{X}_{\gamma_m^{n-1},n}+1.\end{aligned}$$

2. Propagating: Sample the latent variable z_n for the observation y_n from a truncated normal distribution, (17) satisfies :

$$(z^n_{n,m} \mid z^n_{1:n-1,m}, \gamma, y_{1:n}, \mathbf{X}_{1:n}) N_{A_n}(z_n; \widehat{\mu}, \widehat{\sigma}^2),$$

with $A_n = (-\infty, 0]$ for $y_n=0$ and $A_n = (0, +\infty)$ for $y_n=1$, where $N_A(\mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ distribution truncated to A.

3. Rejuvenating:

   a. To sample from (15), cycle though $i=1,\ldots, n$:

$$(z^n_{i,m} \mid z^n_{(i),m}, \gamma^n_m, \mathbf{y}_{1:n}, \mathbf{X}_{1:n},) \propto \begin{cases} N_{(0,+\infty)}(m_i, v_i), y_i=1 \\ N_{(-\infty,0)}(m_i, v_i), y_i=0 \end{cases},$$

   with

$$m_i = \mathbf{x}_{i,\gamma^n_m} B_{\gamma^n_m,n} - w_i(z_i - \mathbf{x}_i B_{\gamma^n_m,n}), v_i = 1+w_i, w_i = h_i/(1-h_i)$$

   and h_i is the i^{th} diagonal element of $\mathbf{H}=\mathbf{X}_{1:n,\gamma^n_m} V_{\gamma^n_m,n} \mathbf{X}'_{1:n,\gamma^n_m}$, $z_{(i)} = (z_1,\ldots, z_{i-1}, z_{i+1},\ldots, z_N)$

   b. To sample $K^{\gamma}_n(\gamma^\star, \mid \mathbf{z}^\star_{1:n}, \gamma)$, we can use the Gibbs sampling kernel or the Metropolis Hasting kernel mentioned in Section 2.1.

Acknowledgments

This research was supported by Grant Number 1R01ES017436-01 from the National Institute of Environmental Health Sciences of the U.S. National Institutes of Health.

References

Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 1993;88:669–679.

Barbieri MM, Berger JO. Optimal predictive model selection. The Annals of Statistics 2004;32:870–897.

Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1995;57:473–484.

Carvalho C, Johannes M, Lopes H, Polson N. Particle learning and smoothing. Statistical Science 2010;25:88–106.

Chopin N. A sequential particle filter method for static models. Biometrika 2002;83:539–552.

Chopin N. Inference and model choice for sequentially ordered hidden markov models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007;69:269–284.

Chung Y, Dunson DB. Nonparametric Bayesian conditional distribution modeling with variable selection. Journal of American Statistical Association 2009;104:1646–1660.

George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association 1993;88:881–889.

George EI, McCulloch RE. Approaches for Bayesian variable selection. Statistica Sinica 1997;7:339–374.

Geweke, J. Variable selection and model comparison in regression. In: Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics 5 – Proceedings of the Fifth Valencia International Meeting; Clarendon Press [Oxford University Press]; 1996. p. 609-620.

Green PJ. Reversible jump markov chain Monte Carlo computation and Bayesian model determination. Biometrika 1995;82:711–732.

Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis 2006:145–168.

Jefferys WH, Berger JO. Ockham's razor and Bayesian analysis. American Scientist 1992;80:64–72.

Liu JS. Monte Carlo Strategies in Scientific Computing. Springer Series in Statistics. 2001

Pitt MK, Shephard N. Filtering via simulation: Auxiliary particle filters. Journal of the American Statistical Association 1999;94:590–599.

Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 1998;92:179–191.

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

Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 2001;1:211–244.

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

Table 1

RMSE of the estimated MIPs, Inclusion percentage and Exclusion percentage based on 50 simulation replicates: PSS with Gibbs kernel (PSSG) and PSS with MH kernel (PSSMH), SSVS with Gibbs (SSVSG) and SSVS with MH kernel (SSVSMH)

		PSSG	PSSMH	SSVSG	SSVSMH
RMSE		0.0112	0.0110	0.0106	0.0110
Ip	α=0.25	0.9987(0.0094)	0.9915(0.0274)	0.9997(0.0014)	0.9902(0.0285)
	α=0.50	0.9945(0.0218)	0.9950(0.0304)	0.9938(0.0312)	0.9943(0.0344)
	α=0.75	0.9748(0.0928)	0.9893(0.0545)	0.9731(0.0785)	0.9865(0.0576)
Ep	α=0.25	0.9700(0.1859)	0.9720(0.1476)	0.9600(0.1979)	0.9800(0.1414)
	α=0.50	0.9945(0.0218)	0.9950(0.0304)	0.9938(0.0312)	0.9943(0.0344)
	α=0.75	0.9848(0.0355)	0.9798(0.0388)	0.9809(0.0387)	0.9781(0.0430)

Table 2

Summaries of the log-marginal likelihoods for the top models in the linear regression case: PSS with Gibbs kernel (PSSG), PSS with MH kernel (PSSMH), SSVS with Gibbs (SSVSG) and SSVS with MH kernel (SSVSMH).

	Median	75%	95%	Maximum
PSSG	−281.4936	−280.8975	−280.5607	−280.5072
PSSMH	−281.4626	−280.8909	−280.5721	−280.5072
SSVSG	−281.5759	−280.9630	−280.6783	−280.6013
SSVSMH	−281.5197	−280.9188	−280.6127	−280.5607

Table 3

The models selected based on different thresholds on the estimated marginal inclusion probabilities for the linear regression case: PSS with Gibbs kernel (PSSG), PSS with MH kernel (PSSMH), SSVS with Gibbs (SSVSG) and SSVS with MH kernel (SSVSMH).

α=0.45	PSSG	2, 3, 5, 6, 7, 9, 14, 15	SSVSG	1, 2, 3, 5, 7, 9, 12, 13, 14, 15
	PSSMH	2, 3, 5, 6, 7, 9, 14, 15	SSVSMH	2, 3, 5, 7, 9, 13, 14, 15
α=0.50	PSSG	2, 3, 5, 7, 9, 14, 15	SSVSG	2, 3, 5, 7, 9, 13, 14, 15
	PSSMH	2, 3, 5, 7, 9, 14, 15	SSVSMH	2, 3, 5, 7, 9, 14, 15
α=0.55	PSSG	3, 5, 7, 9, 14, 15	SSVSG	3, 7, 14, 15
	PSSMH	3, 7, 9, 14, 15	SSVSMH	3, 7, 9, 14, 15

Table 4

Summaries of the log complete data marginal likelihood for the top models selected in the probit case: PSS with Gibbs kernel (PSSG), PSS with MH kernel (PSSMH), MCMC with Gibbs (MCMCG) and MCMC with MH kernel (MCMCMH).

	Median	75%	95%	Maximum
PSSG	−129.8118	−121.9266	−111.6871	−105.0315
PSSMH	−129.7112	−122.3866	−112.4855	−105.0191
MCMCG	−131.5902	−124.8271	−115.2204	−105.0118
MCMCMH	−131.4937	−125.0411	−115.8446	−105.0548

Table 5

The models selected based on different thresholds α on the MIPs in the probit case: PSS with Gibbs kernel (PSSG), PSS with MH kernel (PSSMH), MCMC with Gibbs (MCMCG) and MCMC with MH kernel (MCMCMH).

α=0.45	PSSG	1, 3, 4, 7, 9, 14, 15	MCMCG	1, 2, 3, 4, 7, 9, 10, 14, 15
	PSSMH	1, 3, 4, 7, 9, 10, 14, 15	MCMCMH	1, 2, 3, 4, 5, 7, 9, 10, 14, 15
α=0.50	PSSG	3, 4, 7, 9, 14, 15	MCMCG	2, 3, 4, 7, 9, 14, 15
	PSSMH	3, 4, 7, 9, 14, 15	MCMCMH	1, 3, 4, 7, 9, 14, 15
α=0.55	PSSG	3, 4, 9, 15	MCMCG	3, 4, 9, 14, 15
	PSSMH	3, 4, 9, 15	MCMCMH	3, 4, 9, 15