The New England Journal of Statistics in Data Science Volume 1, 200–207 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS40

Highest Posterior Model Computation and Variable Selection via Simulated Annealing

Arnab Kumar MAITY and Sanjib BASU*

Abstract

Variable selection is widely used in all application areas of data analytics, ranging from optimal selection of genes in large scale micro-array studies, to optimal selection of biomarkers for targeted therapy in cancer genomics to selection of optimal predictors in business analytics. A formal way to perform this selection under the Bayesian approach is to select the model with highest posterior probability. The problem may be thought as an optimization problem over the model space where the objective function is the posterior probability of model. We propose to carry out this optimization using simulated annealing and we illustrate its feasibility in high dimensional problems. By means of various simulation studies, this new approach has been shown to be efficient. Theoretical justifications are provided and applications to high dimensional datasets are discussed. The proposed method is implemented in an R package **sahpm** for general use and is made available on R CRAN.

keywords and phrases: Bayes factor, Highest posterior model, Simulated annealing, Variable selection.

1. INTRODUCTION

Variable selection and the broader problem of model selection remains among the most theoretically and computationally challenging problems, and at the same time, some of the most frequent questions encountered in practice. Jim Berger's contribution in this area are immense and multifaceted, ranging from median probability model [2, 1], g-prior [30], criteria for model choice [4], multiplicity adjustments [33], objective Bayesian methods [6] and many others. In this article, we focus on criterion based Bayesian model selection approaches which include marginal likelihood or Bayes factor based model selection [28], Deviance information criterion (DIC, [37]), log pseudo marginal likelihood (LPML, [22]), or the widely applicable information criterion (WAIC, [41]). The performances of these criteria in model selection is recently compared in [31].

A known challenge to apply these criteria in variable selection problem is the infeasibility to visit all the competing models in the model space even with moderate number of variables, p. [23] noted "for $p > 30$, enumerating all possible models is beyond the reach of modern capability". To emphasize on the difficulty, even with the ultra modern machinery, we refer to [21] where the authors pointed out that a simple binary representation of the full model space with $p = 40$ would occupy 5 terabytes of memory. A possible remedy could be searching the best model over a subset of models such as proposed by [13]. They divided the possible models into few important subsets and then enumerated all the models in those subsets. Another competing variable selection is based on screening the important variables out of all potential covariates the idea of which dates back to [18]. In their work they preselected the significant features according to their marginal correlations with the response variable before applying an embedded method such as lasso [38], which performs variable selection in the process of fitting the model.

We note that any such model selection criterion attempts to favor for a model with lower or the higher criterion values (depending on the criterion). For instance, the highest posterior model (HPM) is the model for which the value of integrated likelihood multiplied by the prior probability, is maximum among competing models in the model space. The idea of HPM is straightforward which makes it a widely accepted criterion for model selection. In addition, as pointed out by [25], HPM enjoys a solid theoretical foundation. Nonetheless, as will be illustrated in this article, the problem of variable selection, using the above argument, may be thought as an maximization problem over the model space, where the objective function is the posterior probability of the models and the optimization is taken place with respect to the models. The optimization approach chosen here is simulated annealing [29].

In HPM based selection one needs to consider three aspects: (1) prior selection, (2) marginal likelihood calculation, and (3) enumeration of model space $\mathcal{M}$. Substantial literature have been devoted to the first two aspects, such as Zellner's g-prior [30] and Laplace approximation with nonlocal priors [27]. The third aspect, namely enumeration of the model space $\mathcal{M}$, as pointed out above, can become infeasible for models with large dimension. Therefore sam-

[*] Corresponding author.

pling from model space or stochastic search of the model space have been suggested in the literature. These include stochastic search by [5], [11], shotgun stochastic search by [24], evolutionary stochastic search by [8], particle stochastic search [34]. On the other hand, recent work of [14] develop Bayesian adaptive sampling (BAS) which is a variant of without replacement sampling according to adaptively updated marginal inclusion probabilities of the variables. More recently, [36] develop a Markov chain Monte Carlo (MCMC) algorithm extending the idea of shotgun stochastic search and screening procedure.

We propose to use simulated annealing for this purpose. Simulated annealing is a stochastic optimization algorithm. It is usually applied to ill-posed problem. The end product of the algorithm is a model which is a collection of explanatory variables among all the variables available, which turns the problem of maximization into a solution of a variable selection problem. There are a number of instances where simulated annealing search has been shown to be effective. For example, [9] used simulated annealing in p-median clustering problem, and [15] used simulated annealing for appropriate portfolio selection. Simulated annealing was also used in feature selection problem, [26] whereas [10] used criteria based on the multiple correlations to carry out the simulated annealing chain.

The rest of the article is organized as follows. In Section 2 we introduce necessary notations, and describe the proposed methodology which we refer to as SA-HPM in this article. In Section 3 we compare the performance of our proposed method with other variable selection techniques in simulation examples. Section 4 illustrates the application of SA-HPM algorithm in two real datasets, one with moderate size of predictors, and the other data is consisted of ultra large number of predictors. Finally we conclude with remarks in Section 5.

2. PROPOSED APPROACH

2.1 Notion of Variable Selection

The focus of this article centers around the linear model where the interest is to explore the linear association between a response variable and the covariates via $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 I)$ where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is the $n \times 1$ response variable, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$, is the $n \times p$ design matrix, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, is $p \times 1$ vector of coefficients. The problem of variable selection can be treated as a model selection problem letting $\mathcal{M} \subseteq \{$all subsets of 1, …, p$\}$ as the model space under consideration. An additional notation of $\boldsymbol{\gamma} \subset \{0,1\}^p$ is introduced to denote $\mathcal{M}_\gamma$, an individual member of $\mathcal{M}$, indexed by the binary vector $\boldsymbol{\gamma}$; while the null model which has no independent variable in the model is denoted by $\mathcal{M}_0$. The stochastic law of representation of $\boldsymbol{Y}$ then depends on $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ via $\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma$ where $\boldsymbol{\gamma}$ is working as a subscript of $\boldsymbol{X}(\boldsymbol{\beta})$ such that $\boldsymbol{x}_j(\boldsymbol{\beta}_j)$ is present in the model whenever $\boldsymbol{\gamma}_j = 1$, $j = 1, \ldots, p$. It follows that there are 2^p models in the model space $\mathcal{M}$, by virtue of which the model space easily becomes large even for moderate p thus precluding to visit all models in the model space.

2.2 Optimization on the Model Space

As discussed before, our focus in this article is on the criterion based variable selection techniques as in any such criteria one must visit all the models in $\mathcal{M}$ to compare and conclude in the favor of a good model. Alternatively, one can find the model having a good, lowest in particular, value of a criterion by performing an optimization on the model space where the optimization can be carried out with respect to the criterion values of the candidate models. More generally, we consider a real valued function $\mathcal{C}$ on $\mathcal{M}$, where $\mathcal{C}$ is the objective function which we want to minimize over $\mathcal{M}$. We recall that, in variable selection setting, any model in the model space can be represented by the binary representation of $\boldsymbol{\gamma}$. So when maximizing any objective function over the model space, the solution must belong to the set of binary numbers $0, 1$. This unique structure of the model space severely limits the choice of optimization methods.

2.3 Simulated Annealing in the Model Space

Because of the special features of the maximization problem, we propose to conduct the maximization process stochastically using the widely known simulated annealing (SA), a stochastic optimization method. In what follows, we provide a brief review of an SA approach; for details, see [7]. We consider the model space $\mathcal{M}$ and define $\mathcal{M}^* \subset \mathcal{M}$ to be the set of global minima of the function $\mathcal{C}$, assumed to be a proper subset of $\mathcal{M}$. For each $i \in \mathcal{M}$, there exists a set $\mathcal{M}(i) \subset \mathcal{M} \setminus \{i\}$, called the set of neighbors of i. In addition, for every i, there exists a collection of positive coefficients $q_{ij}, j \in \mathcal{M}(i)$, such that, $\sum_{j \in \mathcal{M}(i)} q_{ij} = 1$; so $\{q_{ij}\} = Q$ form a transition matrix, elements of which provide the transition probabilities of moving from i to j. It is assumed that $j \in \mathcal{M}(i)$ if and only if $i \in \mathcal{M}(j)$. We also define a nonincreasing function $T : \mathbb{N} \to (0, \infty)$ which is called the cooling schedule. Here $\mathbb{N}$ is the set of positive integers, and $T(t)$ is called the temperature at time t.

Let $\psi(t)$ be a discrete time inhomogeneous Markov chain on the model space $\mathcal{M}$. The search process starts at an initial state $\psi(0) \in \mathcal{M}$. Suppose at time t we arrive at the point i. We then choose a neighbor j of i at random according to probability q_{ij}. Once j is chosen, and if $\mathcal{C}(j) \leq \mathcal{C}(i)$, then $\psi(t+1) = j$ with probability 1; however if $\mathcal{C}(j) > \mathcal{C}(i)$, then $\psi(t+1) = j$ with probability $\exp[-\{\mathcal{C}(j) - \mathcal{C}(i)\}/T(t)]$, otherwise set $\psi(t+1) = i$; this gives raise to the so called Gibbs acceptance probability function. In supplementary material, we provide some technical clarity toward the performance of the proposed method.

[16] established that under regularity conditions, repeating the above steps with gradually reducing the temperature schedule guarantees that $\psi(t)$ converges to the optimal set

M^*, that is, for $k \in \mathbb{N}$ and all $j \in S$

$$\lim_{n\to\infty} \Pr(\psi(n+k) \in \mathcal{M}^* | \psi(k) = i) = \lim_{n\to\infty} \Pr(\mathcal{C}(\psi(n+k)) \tag{2.1}$$

$$= \mathcal{C}^* | \psi(k) = i) = 1 \tag{2.2}$$

where $\mathcal{C}^* = \min_{j\in\mathcal{M}} \mathcal{C}(j)$ and $\mathcal{M}^* = \{i : i \in \mathcal{M}, \mathcal{C}(i) = \mathcal{C}^*\}$.

The conditions for this result to hold are: (i) the probability of moving to j th model from i th model in p steps is positive, that is, $q_{ij}^{(p)} > 0$, $q_{ii} > 0$ for all i, and (iii) Q is irreducible.

2.4 Highest Posterior Model

The Bayesian approach to the variable selection problem is relatively straightforward. We express uncertainty about models by putting a prior distribution on the model $\mathcal{M}_\gamma$, The Bayesian linear model is thus defined as

$$\boldsymbol{Y} \sim \Pr(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}_\gamma), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\mathcal{M}_\gamma), \quad \mathcal{M}_\gamma \sim \Pr(\mathcal{M}_\gamma),$$

where $\pi(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution of parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ under model $\mathcal{M}_\gamma$ and $\Pr(\mathcal{M}_\gamma)$ is the prior on the model $\mathcal{M}_\gamma$. Then posterior distribution of $\boldsymbol{\theta}$ is given by $\pi(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M}_\gamma) = \Pr(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}_\gamma)\pi(\boldsymbol{\theta}|\mathcal{M}_\gamma)/\Pr(\boldsymbol{y}|\mathcal{M}_\gamma)$, where

$$\Pr(\boldsymbol{y}|\mathcal{M}_\gamma) = \int \Pr(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M}_\gamma)\pi(\boldsymbol{\theta}|\mathcal{M}_\gamma)d\boldsymbol{\theta} \tag{2.3}$$

is called the marginal likelihood or integrated likelihood of data $\boldsymbol{y}$ under model $\mathcal{M}_\gamma$. Then the posterior probability of model $\mathcal{M}_\gamma$ can be expressed by

$$\Pr(\mathcal{M}_\gamma|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{y}|\mathcal{M}_\gamma)\Pr(\mathcal{M}_\gamma)}{\sum_{\mathcal{M}_i\in\mathcal{M}} \Pr(\boldsymbol{y}|\mathcal{M}_i)\Pr(\mathcal{M}_i)}. \tag{2.4}$$

The Bayes factor [28, 3] for model $\mathcal{M}_{\gamma_1}$ against model $\mathcal{M}_{\gamma_2}$ is the ratio of their marginal likelihoods, $\mathrm{BF}_{12} = \Pr(\boldsymbol{y}|\mathcal{M}_{\gamma_1})/\Pr(\boldsymbol{y}|\mathcal{M}_{\gamma_2})$. [28] stated that the Bayes factor is a summary of evidence for model $\mathcal{M}_{\gamma_1}$ against model $\mathcal{M}_{\gamma_2}$ and provided a table of cutoffs for interpreting $\log \mathrm{BF}_{12}$. In general, the model with higher log-marginal likelihood is preferable in this model selection criterion.

In modern era, Bayesian inference is typically done by Markov Chain sampling. The computation of Bayes factor from Markov Chain sampling, however, is generally difficult since the Markov Chain methods avoid the computation of the normalizing constant of the posterior and it is precisely this constant that is needed for the marginal likelihood.

The HPM has the highest posterior model probability among all models in the model space, that is, HPM $= \mathrm{argmax}_{\gamma\in\mathcal{M}} \Pr(\mathcal{M}_\gamma|\underline{\boldsymbol{y}})$. Under the notion of a data generating model (or the so-called true model) in the model space it can be shown that the data generating model is often asymptotically equivalent to the highest posterior model. For instance, this can be examined via consistency of posterior model probabilities [20] or via the Bayes factors [32]. [20] examined model consistency for g priors when g is fixed. [30] extended this for mixture of g priors and hyper g priors. [17] proved model consistency for spike and slab type priors. [12] and [32] proved consistency of objective Bayes procedures. On the other hand, [39] and [40] showed Bayes factor consistency for unbalanced ANOVA models and nested designs respectively. Moreover, [27] proved consistency for the true model when non local priors were specified on the parameters. However, they distinguished the true model consistency and pairwise Bayes factor consistency and argued that for large dimensional space pairwise consistency is misleading and hence not much useful.

2.5 The SA-HPM Method

We set $\mathcal{C}$ = negative posterior probability of model $\mathcal{M}_\gamma$ for maximizing the posterior probabilities over the model space applying simulated annealing algorithm. In the SA approach, an appropriately chosen cooling schedule accelerates convergence. When T is very small, the time it takes for the Markov chain $\psi(t)$ to reach equilibrium can be excessive. The main significance of cooling schedule is that, during the beginning of the search process it helps the algorithm to escape from the local modes and then when the search is actually in the neighborhood of the global optimum the algorithm tries to focus in that region by reducing the value of cooling schedule and thereby finding the actual optimum. There is a number of suggestions available in the literature to choose a functional form for cooling schedule.

A transition matrix definition is equally important in an SA algorithm. We define the (i, j) th element of the transition matrix Q as

$$q_{ij} = \frac{\text{posterior probability of } j\text{th model}}{\text{sum of posterior probabilities of neighbors of } i\text{th model}}$$

where j th model $\in$ neighborhood of i th model.

For a given model $\mathcal{M}_\gamma$ we define its neighborhood as $\{\mathcal{M}_{\gamma^0}, \mathcal{M}_{\gamma^{00}}\}$, where γ^0 is such that if $|\gamma_0 - \gamma| = 1$, that is, the model $\mathcal{M}_{\gamma^0}$ can be obtained from model $\mathcal{M}_\gamma$ by either adding or deleting one predictor; $\gamma^{00\prime}1 = \gamma^\prime 1$ and $|\gamma^{00} - \gamma| = 2$, that is, model $\mathcal{M}_{\gamma^{00}}$ can be obtained from model $\mathcal{M}_\gamma$ by swapping one predictor with another.

It is interesting to note that, our selection provides the advantage for getting different region of neighborhood at every step and thus eliminates the possibility of keeping old models in the search region which is the case in [5] and [24]. In this way our approach is different in the sense that the search procedure does not require a complicated and long Markov chain to converge. These ingredients give raise to our proposed stochastic search algorithm called SA-HPM the steps of which are described below. The approach is implemented in R package **sahpm** and is made available on R CRAN.

Step 1: At time t, suppose i = current state of $\gamma(t)$, and set cooling temperature $T(t)$.

Step 2: Choose a neighbor j of i at random according to probability q_{ij}.

Step 3: Once j is chosen, the next state $\gamma(t+1)$ is determined as follows

$$\begin{aligned} &\text{If } \mathcal{C}(j) \leq \mathcal{C}(i), \text{ then } \gamma(t+1) = j \\ &\text{If } \mathcal{C}(j) > \mathcal{C}(i), \text{ then } \gamma(t+1) = j \text{ with probability} \\ &\qquad\qquad \exp[-\{J(j) - J(i)\}/T(t)] \\ &\qquad\qquad \gamma(t+1) = i \text{ otherwise} \end{aligned}$$

If $j \neq i$ and $j \notin S(i)$, then $\Pr[x(t+1) = j | x(t) = i] = 0$.

Step 4: Repeat above steps until convergence.

In practice, to make the computation stable, we suggest to calculate the log of posterior probabilities instead of posterior probabilities and use that as estimates of proposal distributions.

3. OPERATING CHARACTERISTICS IN EMPIRICAL STUDIES

Example 1. In this example we investigate the repeated sampling operating characteristics of complete enumeration based HPM and our proposed SA-HPM method using a cooling temperature $T(t) = 0.9T(t-1)$. Our aim in this example is to see and compare the empirical proprieties of the proposed SA-HPM with those of complete enumeration HPM. As discussed before, the computation for complete enumeration of the model space is feasible only for small p. Hence we focus on those situations whenever complete enumeration HPM computation is feasible. To this end we consider the following simulation models:

1. ($p_{Datagen} = 5$, uncorrelated x's): We simulate data according to the Gaussian linear model $\boldsymbol{Y} \sim N(2 \cdot \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}(D), \sigma^2 I)$ where $\mathbf{1}$ is a column of 1's. We take the data generating model $\mathcal{M}(D)$ to be $\{1,2,3,4,5\}$, $\beta(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$ and $\sigma = (1.5)$. Each row of $\boldsymbol{X}$ is independently generated from $N_p(0, \Sigma_X)$ where $\Sigma_X = I$ is taken to be isotropic.
2. ($p_{Datagen} = 5$, correlated x's) The rows of X are generated so that $cor(x_i, x_j) = \rho$ for all $i \neq j$. We take $\rho = 0.5$.
3. ($p_{Datagen} = 5$, autoregressive correlated X's): The rows of X are generated so that $var(x_i) = 1$, and $cor(x_i, x_j) = \rho^{|i-j|}$ for $i \neq j$. We take $\rho = 0.5$.

For each setting we consider $p = 15$ and $p = 20$, two sample sizes $n = 100$, and $n = 1000$, and 100 replicated datasets for each combination. We use $g = \max(n, p^2)$ in the g prior [19]. Table 1 summarizes the simulation result. We notice that both methods report low false discovery rate and false nondiscovery rate. Furthermore, both HPM and SA-HPM perform satisfactorily in terms of recovering the data generating model. For instance, when $n = 100$, $p = 15$, and the variance covariance matrix of the design matrix is isotropic, the proportions of time the data generating model got identified by SA-HPM and HPM are 0.84 and 0.86 respectively. Similar performance is evident for other settings however is slightly worse when $p = 20$. Nevertheless, the important finding to note here is that the performance of the SA-HPM method is comparable to that of the complete enumeration HPM.

Table 1. $\mathcal{M}(D)$ is the proportion of times the data generating model is selected. FDR is the false discovery rate (= FP/(TP+FP)) and FNDR is the false nondiscovery rate (= FN/(TP+FN)), both averaged over replications. Here TP, FP and FN are "True Positive", "False Positive" and "False Negative" counts respectively. Results are based on 100 replications.

	SA-HPM			HPM		
	$\mathcal{M}(D)$	FDR	FNDR	$\mathcal{M}(D)$	FDR	FNDR
	$n = 100$, $p = 15$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$					
$\rho = 0$	0.84	0.02	0.00	0.86	0.02	0.00
$\rho = 0.5$	0.82	0.02	0.00	0.86	0.02	0.00
AR	0.87	0.01	0.00	0.89	0.01	0.00
	$n = 1000$, $p = 15$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$					
$\rho = 0$	0.84	0.02	0.00	0.94	0.01	0.00
$\rho = 0.5$	0.83	0.02	0.00	0.86	0.02	0.00
AR	0.85	0.02	0.00	0.96	0.00	0.00
	$n = 100$, $p = 20$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$					
$\rho = 0$	0.74	0.02	0.00	0.77	0.02	0.00
$\rho = 0.5$	0.83	0.01	0.00	0.83	0.01	0.00
AR	0.77	0.02	0.00	0.87	0.01	0.00
	$n = 1000$, $p = 20$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$					
$\rho = 0$	0.75	0.02	0.00	0.88	0.01	0.00
$\rho = 0.5$	0.84	0.01	0.00	0.83	0.01	0.00
AR	0.78	0.02	0.00	0.83	0.01	0.00

Example 2. In this example we investigate and compare the performance of SA-HPM method to the nonlocal prior based selection [27] whose theoretical and numerical performances are recently considered in [36] and we use its stochastic search implementation is in the R package **BayesS5** [35] in its default setting. We consider similar settings of uncorrelated, equi-correlated, and auto-correlated design matrices from covariate space as in the previous example. We set $n = 100$ and vary $p = 30, 200$, and 1000. In addition, we consider a special correlated design matrix where rows of $\boldsymbol{X}$ are generated so that $var(\boldsymbol{x}_j) = 1$, $\text{cor}(\boldsymbol{x}_4, \boldsymbol{x}_j) = \rho^{1/2}, j \neq 4$, and $\text{cor}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho$ for all other $i \neq j, \rho = 0.5$. We take the data generating model $\mathcal{M}(D)$ to be 1, 2, 3, 4, and $\boldsymbol{\beta}(D) = (5, 5, 5, -15)$. We note that, in this way, $\boldsymbol{x}_4$ is uncorrelated with the response $\boldsymbol{Y}$ [18].

As in the previous example, we consider a g prior on the regression coefficients for our proposed SA-HPM method. Additionally, motivated by the beautiful properties of nonlocal prior [27], we specify piMOM prior as a representative of

Table 2. $\mathcal{M}(D)$ is the proportion of times the data generating model is selected. FDR is the false discovery rate (= FP/(TP+FP)) and FNDR is the false nondiscovery rate (= FN/(TP+FN)), both averaged over replications. Here TP, FP and FN are “True Positive”, “False Positive” and “False Negative” counts respectively. Results are based on 100 replications.

	SA-HPM-g			SA-HPM-piMOM			BayesS5		
	$\mathcal{M}(D)$	FDR	FNDR	$\mathcal{M}(D)$	FDR	FNDR	$\mathcal{M}(D)$	FDR	FNDR
	$n = 100$, $p = 30$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$								
$\rho = 0$	0.87	0.02	0.00	0.99	0.00	0.00	0.92	0.01	0.00
$\rho = 0.5$	0.78	0.04	0.00	0.95	0.01	0.00	0.95	0.01	0.00
AR	0.82	0.03	0.00	1.00	0.00	0.00	0.97	0.00	0.00
	$n = 100$, $p = 200$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$								
$\rho = 0$	0.87	0.02	0.00	0.99	0.00	0.00	0.92	0.01	0.00
$\rho = 0.5$	0.81	0.02	0.00	0.94	0.00	0.00	0.95	0.01	0.00
AR	0.80	0.03	0.00	0.99	0.00	0.00	0.99	0.00	0.00
	$n = 100$, $p = 1000$, $\boldsymbol{\beta}(D) = (1.5, -1.5, 1.5, -1.5, 1.5)$								
$\rho = 0$	0.87	0.03	0.00	1.00	0.00	0.00	0.99	0.00	0.00
$\rho = 0.5$	0.71	0.02	0.00	0.70	0.06	0.00	1.00	0.00	0.00
AR	0.80	0.03	0.00	0.94	0.00	0.00	1.00	0.00	0.00
	$n = 100$, $\boldsymbol{\beta}(D) = (5, 5, 5, -15)$, cor$(\boldsymbol{x}_4, \boldsymbol{x}_j) = \rho^{1/2}, \rho = 0.5, j \neq 4$								
$p = 30$	0.79	0.04	0.00	0.98	0.02	0.00	0.97	0.01	0.00
$p = 200$	0.70	0.18	0.00	0.89	0.09	0.00	0.75	0.18	0.00
$p = 1000$	0.52	0.36	0.00	0.65	0.31	0.00	0.39	0.58	0.00

the class of nonlocal priors and use Laplace approximation to obtain marginal likelihood. We refer to these two procedures as SA-HPM-g and SA-HPM-piMOM respectively. We present the simulation result in Table 2 and notice that SA-HPM with piMOM prior outperforms the SA-HPM with g prior, particularly in high dimensional settings. Furthermore, we observe similar performances of SA-HPM-piMOM prior and BayesS5 method; however, when $\boldsymbol{x}_4$ is uncorrelated with the response then the performance of BayesS5 deteriorates.

4. APPLICATION IN HIGH-DIMENSIONAL SELECTION SETTINGS

4.1 Ozone35 Data, Moderate p

The ozone dataset has been considered in the literature frequently [5, 11] and consists of daily measurements of atmospheric ozone concentration (maximum one hour average) and eight meteorological quantities for 330 days of 1976 in the Los Angeles Basin. Among them one temperature predictor was dropped from the analysis due to the potential multicollinearity with another temperature variable. The Ozone35 data was then curated by considering the main effects, the second order effects, their first order interactions [21], which gives raise to a total of $p = 35$ covariates. Mainly for comparison purpose we make use of the $n = 178$ observations which were used in the analysis of [21]. The description of the predictor variables and the response variable is provided in Table 3. [21] illustrated that the posterior probability of the median probability model (MPM [2]) is 23 times lower than that of the highest posterior model. We considered a g-prior as in [21].

[21] considered a complete enumeration of this large model space using distributed computing over an extended time and reported the complete enumeration HPM to be the model (7, 10, 23, 26, 29). The Bayes factor of HPM and MPM, against $\mathcal{M}_0$ are reported in Table 4. Our main contribution is that, the proposed SA-HPM method is able to recover the HPM 95 times out of 100 repetitions after a burn-in of 50 iterations in the stochastic chain. In particular, we note that, the SA-HPM method is extremely useful even for large model spaces.

4.2 Polymerase Chain Reaction Data, Ultra Large p

In this example we consider gene expression data on 31 female mice and 29 male mice. A number of psychological phenotypes, including numbers of stearoyl-CoA desaturase 1 (SCD1), glycerol-3-phosphate acyltransferase (GPAT) and phos- phoenopyruvate carboxykinase (PEPCK), were measured by quantitative real-time RT-PCR, along with 22,575 gene expression values. The resulting data set is publicly available at http://www.ncbi.nlm.nih.gov/geo (accession number GSE3330). Following [35] we restrict ourselves into the consideration of the SCD1 response only.

Due to ultra large high-dimensional nature of this dataset it is beyond the reach of the ultra modern machinery to enumerate all the models in the model space. Hence, in order to find the highest posterior model it is necessary to make use of a model space search technique such as the SA-HPM method developed here. We employ our algorithm in this dataset to find the HPM model. When utilizing SA-HPM we omit the swapping step to minimize exploring the model

Table 3. Description of the Ozone35 dataset variables.

Variable	Description
y	Response = Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA
x_1	500-millibar pressure height (m) measured at Vandenberg AFB
x_2	Wind speed (mph) at Los Angeles International Airport (LAX)
x_3	Humidity (%) at LAX
x_4	Temperature (F) measured at Sandburg, CA
x_5	Inversion base height (feet) at LAX
x_6	Pressure gradient (mm Hg) from LAX to Daggett, CA
x_7	Visibility (miles) measured at LAX

Table 4. Table with two rows indicating HPM and MPM respectively. The last two columns provide Bayes factor against the null model and log of that respectively.

Serial No	Model	Bayes Factor	log(Bayes Factor)
HPM	7 10 23 26 29	1.02E+47	108.2364944
MPM	21 22 23 29	4.34E+45	105.0834851

space due to the ultra large size of that. We report our findings in Table 5 from which it can be noted that the resulting HPM is a sparse model with three variables when SA-HPM with g prior is fitted. Similarly, SA-HPM with piMOM prior discovers another sparse model with two predictors. It is interesting to note that one of them coincides with the model projected by the maximum aposteriori (MAP) estimate of the BayesS5 method. We notice that BayesS5 also results in a parsimonious model with two predictor variables.

Table 5. Resulting models in Polymerase Chain Reaction Data.

Method	Model
SA-HPM-g	5905 8422 12999
SA-HPM-piMOM	296 5510
BayesS5	296 7351

5. CONCLUSION

We note that, our approach is distinguishable from the many traditional approaches in this area in terms of the fact that the methodology developed in this work does not aim to recover the data generating model rather our effort focuses on finding the highest posterior model which is often perceived to have good properties. If highest posterior model does not coincide with the data-generating model, our proposed SA-HPM method is still able to recover the HPM without finding the data-generating one. In a real world data analysis the data generating model or the so called “true model” is not known and hence our approach is useful to consider.

As a summary, our research strengthens the classical idea of assessing a model by its posterior probability. According to [21], a large volume of near future research in Bayesian literature of variable selection will involve sampling and stochastic search. Furthermore, [23] noted that good models can be obtained by exploring the posterior summary of the models. Nonetheless, the highest posterior model, a posterior summary, is widely known to have excellent properties. Our research, thus, provides a simple, efficient, quick, and feasible way toward this direction of variable selection.

SUPPLEMENTARY MATERIAL

The R package **sahpm** for the method SA-HPM is available on R CRAN. Further mathematical discussion on the convergence of this method is given in a separate supplementary material.

ACKNOWLEDGEMENTS

The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of NIEHS. The computational work in this manuscript utilized resources of the Center for Research Computing and Data at Northern Illinois University.

FUNDING

Sanjib Basu’s research was partially supported by award R01-ES028790 from the National Institute of Environmental Health Sciences.

Accepted 30 May 2023

REFERENCES

[1] Barbieri, M. M., Berger, J. O., George, E. I. and Ročková, V. (2021). The median probability model and correlated variables. *Bayesian Analysis* **16**(4) 1085–1112. https://doi.org/10.1214/20-BA1249. MR4381128

[2] Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The annals of statistics* **32**(3) 870–897. https://doi.org/10.1214/009053604000000238. MR2065192

[3] Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association* **98**(461) 224–235. https://doi.org/10.1198/016214503388619472. MR1965688

[4] Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics* **40**(3) 1550–1577. https://doi.org/10.1214/12-AOS1013. MR3015035

[5] Berger, J. O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**(1) 3–15. https://doi.org/10.1111/j.1467-9574.2005.00275.x. MR2137378

[6] Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J. and Pericchi, L. (2001). *Objective Bayesian methods for model selection: Introduction and comparison. Lecture Notes-Monograph Series* 135–207. https://doi.org/10.1214/lnms/1215540968. MR2000753

[7] Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statistical science* **8**(1) 10–15. MR1194437

[8] Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis* **5**(3) 583–618. https://doi.org/10.1214/10-BA523. MR2719668

[9] Brusco, M. J. and Köhn, H. q. F. (2009). Exemplar-based clustering via simulated annealing. *Psychometrika* **74**(3) 457–475. https://doi.org/10.1007/s11336-009-9115-2. MR2551671

[10] Cadima, J., Cerdeira, J. O. and Minhoto, M. (2004). Computational aspects of algorithms for variable selection in the context of principal components. *Computational Statistics & Data Analysis* **47**(2) 225–236. https://doi.org/10.1016/j.csda.2003.11.001. MR2101498

[11] Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101**(473) 157–167. https://doi.org/10.1198/016214505000000646. MR2268035

[12] Casella, G., Girón, F. J., Martínez, M. L. and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Annals of Statistics* **37**(3) 1207–1228. https://doi.org/10.1214/08-AOS606. MR2509072

[13] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**(3) 759–771. https://doi.org/10.1093/biomet/asn034. MR2443189

[14] Clyde, M. A., Ghosh, J. and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics* **20**(1) 80–101. https://doi.org/10.1198/jcgs.2010.09049. MR2816539

[15] Crama, Y. and Schyns, M. (2003). Simulated annealing for complex portfolio selection problems. *European Journal of Operational Research* **150**(3) 546–571.

[16] Cruz, J. R. and Dorea, C. C. Y. (1998). Simple conditions for the convergence of simulated annealing type algorithms. *Journal of Applied Probability* **35**(4) 885–892. https://doi.org/10.1239/jap/1032438383. MR1671238

[17] Dey, T., Ishwaran, H. and Rao, J. S. (2008). An in-depth look at highest posterior model selection. *Econometric Theory* **24**(2) 377–403. https://doi.org/10.1017/S026646660808016X. MR2391616

[18] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5) 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x. MR2530322

[19] Fernandez, C., Ley, E. and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**(2) 381–427. https://doi.org/10.1016/S0304-4076(00)00076-2. MR1820410

[20] Fernandez, C., Ley, E. and Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**(5) 563–576.

[21] Garcia-Donato, G. and Martinez-Beneito, M. A. (2013). On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association* **108**(501) 340–352. https://doi.org/10.1080/01621459.2012.742443. MR3174624

[22] Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**(365) 153–160. MR0529531

[23] Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association* **110**(509) 435–448. https://doi.org/10.1080/01621459.2014.993077. MR3338514

[24] Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102**(478) 507–516. https://doi.org/10.1198/016214507000000121. MR2370849

[25] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**(4)382–401. https://doi.org/10.1214/ss/1009212519. MR1765176

[26] Jeong, I. q. S., Kim, H. q. K., Kim, T. q. H., Lee, D. H., Kim, K. J. and Kang, S. q. H. (2018). A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks. *Software Networking* **2018**(1) 173–190.

[27] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107**(498) 649–660. https://doi.org/10.1080/01621459.2012.682536. MR2980074

[28] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**(430) 773–795. https://doi.org/10.1080/01621459.1995.10476572. MR3363402

[29] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**(4598) 671–680. https://doi.org/10.1126/science.220.4598.671. MR0702485

[30] Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103**(481) 410–423. https://doi.org/10.1198/016214507000001337. MR2420243

[31] Maity, A. K., Basu, S. and Ghosh, S. (2021). Bayesian criterion-based variable selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **70**(4) 835–857. https://doi.org/10.1111/rssc.12488. MR4318011

[32] Moreno, E., Girón, F. J. and Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *Annals of Statistics* **38**(4) 1937–1952. https://doi.org/10.1214/09-AOS754. MR2676879

[33] Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38**(5) 2587–2619. https://doi.org/10.1214/10-AOS792. MR2722450

[34] Shi, M. and Dunson, D. B. (2011). Bayesian variable selection via particle stochastic search. *Statistics & probability letters* **81**(2) 283–291. https://doi.org/10.1016/j.spl.2010.10.011. MR2764295

[35] Shin, M. and Tian, R. (2017). BayesS5: Bayesian Variable Selection Using Simplified Shotgun Stochastic Search with Screening (S5). R package version 1.30. https://CRAN.R-project.org/package=BayesS5.

[36] Shin, M., Bhattacharya, A. and Johnson, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica* **28**(2) 1053. MR3791100

[37] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4) 583–639. https://doi.org/10.1111/1467-9868.00353. MR1979380

[38] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1) 267–288. MR1379242

[39] Wang, M. and Sun, X. (2013). Bayes factor consistency for unbalanced ANOVA models. *Statistics* **47**(5) 1104–1115. https://doi.org/10.1080/02331888.2012.694445. MR3175737

[40] Wang, M. and Sun, X. (2014). Bayes factor consistency for nested linear models with a growing number of parameters. *Journal of Statistical Planning and Inference* **147** 95–105. https://doi.org/10.1016/j.jspi.2013.11.001. MR3151848

[41] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**(Dec) 3571–3594. MR2756194

Arnab Kumar Maity. Pfizer, USA. E-mail address: Arnab.Maity@pfizer.com

Sanjib Basu. University of Illinois Chicago, USA. E-mail address: sbasu@uic.edu