Bayesian Adaptive Sampling for Variable Selection and Model Averaging

Merlise Clyde*, Joyee Ghosh† and Michael Littman‡

Abstract

For the problem of model choice in linear regression, we introduce a Bayesian adaptive sampling algorithm (BAS), that samples models without replacement from the space of models. For problems that permit enumeration of all models BAS is guaranteed to enumerate the model space in 2^p iterations where p is the number of potential variables under consideration. For larger problems where sampling is required, we provide conditions under which BAS provides perfect samples without replacement. When the sampling probabilities in the algorithm are the marginal variable inclusion probabilities, BAS may be viewed as sampling models "near" the median probability model of Barbieri and Berger. As marginal inclusion probabilities are not known in advance we discuss several strategies to estimate adaptively the marginal inclusion probabilities within BAS. We illustrate the performance of the algorithm using simulated and real data and show that BAS can outperform Markov chain Monte Carlo methods. The algorithm is implemented in the `R` package `BAS` available at CRAN.

Key words: Bayesian model averaging; Inclusion probability; Markov chain Monte Carlo; Median probability model; Model uncertainty; Sampling without replacement

1 INTRODUCTION

We consider the problem of model uncertainty in linear regression with p potential predictors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$. In this setting, models $\mathcal{M}_{\boldsymbol{\gamma}}$ may be represented as a vector of binary variables $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)^T \in \{0,1\}^p \equiv \Gamma$ where γ_j is an indicator of whether $\boldsymbol{x}_j$ is included as a column

*Merlise Clyde is Associate Professor of Statistics, Duke University, Durham, NC 27705. Email `clyde@stat.duke.edu`

†Joyee Ghosh is Postdoctoral Fellow, Department of Biostatistics, The University of North Carolina, Chapel Hill, NC 27599. Email `jghosh@bios.unc.edu`

‡Michal Littman is Professor, Computer Sciences, Rutgers University, Piscataway, NJ 08854. Email `mlittman@cs.rutgers.edu`

in the $n \times p_{\boldsymbol{\gamma}}$ design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ under model $\mathcal{M}_{\boldsymbol{\gamma}}$. The normal linear model conditional on $\mathcal{M}_{\boldsymbol{\gamma}}$ is expressed as

$$\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \mathcal{M}_{\gamma} \sim \mathsf{N}(\mathbf{1}_n \alpha + \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{I}_n/\phi) \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots Y_n)'$, $\mathbf{1}_n$ denotes a vector of ones of length n, α is the intercept, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ represents the regression coefficients and ϕ is the precision (the inverse of the error variance) with $\mathbf{I}_n$ denoting the $n \times n$ identity matrix.

A key component in the Bayesian formulation of the model choice problem is the posterior distribution over models given by

$$p(\mathcal{M}_{\boldsymbol{\gamma}} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) p(\mathcal{M}_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) p(\mathcal{M}_{\boldsymbol{\gamma}})} \tag{2}$$

where $p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \mathcal{M}_{\boldsymbol{\gamma}}) p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \mathcal{M}_{\boldsymbol{\gamma}}) d\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ is proportional to the marginal likelihood of $\mathcal{M}_{\boldsymbol{\gamma}}$ obtained by integrating the joint likelihood with respect to the prior distribution over all parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$ given $\mathcal{M}_{\boldsymbol{\gamma}}$ and $p(\mathcal{M}_{\boldsymbol{\gamma}})$ is the prior probability of the model.

The joint posterior distribution over models and model specific parameters provides the basis for decisions regarding model choice. Bayesian model averaging (BMA) utilizes the full joint posterior distribution and incorporates model uncertainty in posterior inferences, see Hoeting et al. (1999); Clyde and George (2004) for overviews of BMA. For a specific quantity of interest Δ, the posterior distribution under BMA is

$$p(\Delta \mid \mathbf{Y}) = \sum_{\boldsymbol{\gamma} \in \Gamma} p(\Delta \mid \mathcal{M}_{\gamma}, \mathbf{Y}) p(\mathcal{M}_{\gamma} \mid \mathbf{Y}) \tag{3}$$

with model averaged expectations of the form

$$\mathsf{E}[\Delta \mid \mathbf{Y}] = \sum_{\boldsymbol{\gamma} \in \Gamma} \mathsf{E}[\Delta \mid \mathcal{M}_{\gamma}, \mathbf{Y}] p(\mathcal{M}_{\gamma} \mid \mathbf{Y}). \tag{4}$$

When it is necessary to report a single model, a common strategy is to select the highest posterior probability model, $\boldsymbol{\gamma}_{\mathsf{HPM}}$ which corresponds to maximizing a 0-1 utility for a correct selection. Other variable selection procedures can be formally motivated by decision theoretic

considerations, where the optimal model, $\boldsymbol{\gamma}^*$, maximizes posterior expected utility. If the goal is to select the best predictive model under squared error loss, Barbieri and Berger (2004) show that for a sequence of nested models, the median probability model $\boldsymbol{\gamma}_{\mathsf{MPM}}$:

$$(\gamma_j)_{\mathrm{MPM}} \equiv \begin{cases} 1 & \text{if} \quad \mathsf{P}(\gamma_j = 1 \mid \mathbf{Y}) \equiv \pi_j \geq 1/2 \\ 0 & \quad \text{otherwise} \end{cases} \tag{5}$$

is optimal. While the median probability model is not optimal under squared error loss for general design matrices in regression, Barbieri and Berger (2004) suggest that in practice the median probability model $\boldsymbol{\gamma}_{\mathsf{MPM}}$ is often preferable to the highest posterior probability model $\boldsymbol{\gamma}_{\mathsf{HPM}}$. The median probability model is also the centroid estimator, which minimizes a Hamming loss function (Carvalho and Lawrence 2008). Carvalho and Lawrence (2008) argue that the centroid estimator does a better job of capturing the character of the distribution than the highest probability model. For general designs, the optimal model for prediction under squared error loss is the model whose predictions are closest to those under BMA; when multicollinearity is present this model and the median probability model may be quite different.

When the number of variables p is greater than 25-30, enumeration of all possible models in Γ is generally intractable, and sampling or search methods are necessary, irrespective of whether the goal is to determine an optimal model, or to make inferences and predictions based on BMA. The `BMA` package in `R` utilizes deterministic sampling using the leaps and bounds algorithm (Furnival and Wilson 1974; Hoeting et al. 1999) which attempts to find the q "best" models of a given dimension. Because leaps considers all dimensions, this can be inefficient in large problems. Stochastic search variable selection, SSVS, (George and McCulloch 1997) and the related MC^3 algorithm (Raftery et al. 1997) are popular Markov Chain Monte Carlo (MCMC) algorithms that can be viewed as providing a (dependent) stochastic sample of models from the posterior distribution on Γ. While easy to implement (SSVS is a Gibbs sampler, while MC^3 is a random-walk Metropolis sampler), these early algorithms may mix poorly when covariates are highly correlated; more advanced algorithms for the variable selection problem that utilize other proposals include adaptive MCMC (Nott

and Kohn 2005), Swendsen-Wang (Nott and Green 2004) and Evolutionary Monte Carlo (Liang and Wong 2000; Wilson et al. 2010; Bottolo and Richardson 2008).

Historically, the conjugate Normal-Gamma family of prior distributions has received widespread attention for model choice in linear models (Raftery et al. 1997; Smith and Kohn 1996; George and McCulloch 1997) as marginal likelihoods can be evaluated analytically. Of these, Zellner's g-prior (Zellner 1986) remains perhaps the most popular conventional prior distribution with marginal likelihoods that may be expressed as a simple function of the model R^2. Mixtures of Zellner's g-prior, such as the Zellner-Siow Cauchy prior (Zellner and Siow 1980) or the hyper-g prior of Liang et al. (2008) retain the computational simplicity of the original g-prior, but resolve many of the inconsistencies that arise from using a fixed g. For such mixtures, marginal likelihoods may be obtained as a one dimensional integral with respect to a prior on g, which is available in closed form for the hyper-g prior or may be approximated via a one dimensional numerical integration or a Laplace approximation in the case of the Zellner-Siow prior (see the review article by Liang et al. (2008) for computational and theoretical details).

MCMC methods are often used as a model search strategy for identifying high probability models for selection or model averaging. When marginal likelihoods are available, these quantities are often used in place of the MCMC model frequencies for ranking or selecting models (George and McCulloch 1997) or model averaging over a subset of models (Raftery et al. 1997) as they provide exact Bayes factors for comparing any two models or exact (conditional) model probabilities for a restricted set of models. In this context, Clyde (1999) suggested that sampling models without replacement from the model space may be a more efficient strategy than re-sampling models if the Monte Carlo frequencies of model visits are not utilized in estimation. In this paper, we construct a novel algorithm that samples models subject to the constraint of producing no duplicates. We give conditions under which this provides perfect samples without replacement from the posterior distribution over models. For models where these conditions do not hold, we develop a Bayesian adaptive sampling (BAS) algorithm which provides sequential learning of the marginal inclusion probabilities, while sampling without replacement.

The paper is arranged in the following manner. In Section 2, we discuss sampling without replacement from the space of models and show how to construct an adaptive, stochastic sampling without replacement algorithm. In Section 3, we discuss several strategies to provide sampling probabilities which are used to initialize the algorithm. In Section 4, we compare BAS to MCMC in a simulation study and show that BAS can outperform MCMC algorithms on several criteria. Sections 5 and 6 illustrate the method in two real data sets: the U.S. crime data, where enumeration is feasible, and the moderate dimension protein construct data (Clyde et al. 1996) where exhaustive search is not possible. In Section 7 we conclude with recommendations and a discussion of possible extensions.

2 SAMPLING WITHOUT REPLACEMENT

Sampling without replacement from a finite population such as the space of models Γ can be implemented in a variety of ways. The most straightforward sampling mechanism is to draw a simple random sample without replacement (SRSWOR) of size T from Γ, assigning equal probabilities to all models in Γ. Although conceptually simple, it is quite likely that many high posterior probability models will not be included in the sample of models unless T is large relative to $|\Gamma|$.

An improvement over SRSWOR can be achieved by drawing a random sample without replacement of size T from Γ where the probability of sampling a model depends on a measure of the model's "importance" or "size". In such probability proportional to size (PPS) sampling, one constructs "size" variables for all models in the model space (ideally such that they are highly correlated with product of the marginal likelihood and prior probability of each model), and then samples models with probabilities proportional to their "size". After a model is sampled, its size contribution is subtracted off and the size variables of the remaining models are re-normalized, and the next model is selected using the new size variables. A complicating factor in implementing PPS sampling is that the sampling frame which specifies all models and "size" variables cannot be listed exhaustively prior to sampling, as this is of the same computational complexity as enumerating the model space. In what follows, we present a novel way to sample from the model space via PPS sampling that bypasses the

need to construct the sampling frame.

2.1 Sampling without Replacement on Binary Trees

The Bayesian Adaptive Sampling (BAS) algorithm is designed to sample models without replacement such that the probability of a model being sampled is proportional to some probability mass function $f(\boldsymbol{\gamma})$ with known normalizing constant. In BAS, the model space Γ is represented by a binary tree with γ_1 at the top node followed by $\gamma_2, \ldots \gamma_p$ respectively, as shown in Figure 1 with $p = 3$. Two branches arise from each node j, with the left and right branches given by γ_j equal to 0 and 1 respectively. Each model in Γ is represented as a unique path among 2^p possible paths in the binary tree. This form facilitates computing the re-normalized probabilities without listing all 2^p models in advance and permits direct sampling without replacement from the re-normalized probabilities on the tree. We illustrate the algorithm first in the case with $p = 3$ before giving a more formal description of the algorithm.

2.2 Illustration

Suppose that we want to take samples without replacement from a product Bernoulli distribution of the form $f(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \rho_j^{\gamma_j}(1-\rho_j)^{1-\gamma_j}$. We may draw the first model $\boldsymbol{\gamma}^{(1)}$ directly by generating each γ_j as an independent Bernoulli with probability ρ_j. In order to sample without replacement, every time a new model is sampled, one needs to account for its mass by subtracting off its probability from $f(\boldsymbol{\gamma})$ to ensure that there is no duplication and then draw a new model from the re-normalized probability distribution. Figure 1 provides an illustration of sampling without replacement for $p = 3$ where the model space contains $2^p = 8$ models and $\rho_1 = 3/4$, $\rho_2 = 1/2$ and $\rho_3 = 1/4$.

Figure 1 (a) shows the first model $\boldsymbol{\gamma}^{(1)} = (0, 0, 0)$ being sampled (the branches with a solid line). Next Figure 1 (b) shows that the mass of the sampled model has been subtracted off leading to the re-normalized distribution on the tree. Note only the sampling probabilities along the path of the sampled model require updating to $24/29, 4/5$ and 1 for $j = 1, 2$, and 3 respectively, while the probabilities for the other branches remain unchanged from their

initial values. The sequence of plots show that as all models below a node are sampled, the branch receives zero probability. The last plot at $t = 8$ (Figure 1 (h)), shows that the last model $\boldsymbol{\gamma}^{(8)} = (0, 1, 0)$ is sampled with probability 1, thus completely enumerating the model space. We have shown the probabilities for all the branches at each iteration for illustrative purposes only; in the implementation they are not calculated/updated until a node is sampled.

We now give a general description of how the updating of the distribution is achieved.

2.3 Algorithm for Sampling without Replacement on Binary Trees

Any probability mass function may be written as a sequence of conditional and marginal distributions of the form

$$f(\boldsymbol{\gamma}) = \prod_{j=1}^{p} f(\gamma_j \mid \boldsymbol{\gamma}_{<j}) \tag{6}$$

where the notation $\boldsymbol{\gamma}_{<j}$ indicates the subset of inclusion indicators $\{\gamma_k\}$ for $k < j$. Similarly we will let $\boldsymbol{\gamma}_{\geq j} = \{\gamma_k\}$ for $k \geq j$. For $j = 1$, $f(\gamma_1 \mid \boldsymbol{\gamma}_{<1}) \equiv f(\gamma_1)$ is the marginal distribution of γ_1. Because the γ_j are binary, we may re-express (6) as

$$f(\boldsymbol{\gamma} \mid \boldsymbol{\rho}) = \prod_{j=1}^{p} (\rho_{j|<j})^{\gamma_j} (1 - \rho_{j|<j})^{1-\gamma_j} \tag{7}$$

where $\rho_{j|<j} \equiv f(\gamma_j = 1 \mid \boldsymbol{\gamma}_{<j})$ and $\boldsymbol{\rho}$ is the collection of all $\{\rho_{j|<j}\}$. After sampling a model from (7), the distribution on the remaining models is of the same form as (7), but with a new $\boldsymbol{\rho}$. The updating of the sampling probabilities, $\boldsymbol{\rho}$ is given by the following theorem:

Theorem 1. *Let $\mathbb{S}_t$ denote the set of sampled models at time t and let Γ_t represent the remaining unsampled models such that $\Gamma_t \cup \mathbb{S}_t = \Gamma$ and $\Gamma_t \cap \mathbb{S}_t = \emptyset$ with $\mathbb{S}_0 \equiv \emptyset$. Let $f(\boldsymbol{\gamma} \mid \boldsymbol{\rho}^{(t)})$ denote the current probability mass function of the form given by (7), where $\boldsymbol{\rho}^{(t)}$ is the set of probabilities $\{\rho^{(t)}_{j|<j}\}$ such that $f(\gamma \mid \boldsymbol{\rho}^{(t)})$ assigns probability one to Γ_t. To begin, we initialize $\boldsymbol{\rho}^{(0)} = \boldsymbol{\rho}$.*

For $t = 1, \ldots, T$, let $\boldsymbol{\gamma}^{(t)}$ denote the model sampled at the t^{th} step where

$$\boldsymbol{\gamma}_j^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)} \sim \mathsf{Ber}\left(\rho_{j|<j}^{(t-1)}\right)$$

and set $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{\boldsymbol{\gamma}^{(t)}\}$. Let

$$f(\boldsymbol{\gamma}_{\geq j}^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)}) = \prod_{k=j}^{p} \left(\rho_{k|<k}^{(t-1)}\right)^{\gamma_k^{(t)}} \left(1 - \rho_{k|<k}^{(t-1)}\right)^{1-\gamma_k^{(t)}} \tag{8}$$

denote the sampling probability of the nodes $\gamma_k^{(t)}$ for $k \geq j$ of the current branch $\boldsymbol{\gamma}^{(t)}$. For $j = 1, \ldots, p$, the conditional probabilities $\rho_{j|<j}^{(t-1)}$ for the branch $\boldsymbol{\gamma}^{(t)}$ are updated to

$$\rho_{j|<j}^{(t)} = \frac{\rho_{j|<j}^{(t-1)} - f(\boldsymbol{\gamma}_{\geq j}^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)})\gamma_j^{(t)}}{1 - f(\boldsymbol{\gamma}_{\geq j}^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)})} \tag{9}$$

while for all other branches

$$\rho_{j|<j}^{(t)} = \rho_{j|<j}^{(t-1)}. \tag{10}$$

Then $f(\boldsymbol{\gamma} \mid \boldsymbol{\rho}^{(t)})$ assigns zero mass to any previously sampled models $\boldsymbol{\gamma} \in \mathcal{S}_t$ and probability one to the remaining space of models $\Gamma_t = \Gamma_{t-1} - \{\boldsymbol{\gamma}^{(t)}\}$.

The proof is given in the Supplemental Materials. The key idea of the proof and algorithm is showing that the $\rho_{j|<j}^{(t)}$ are the conditional inclusion probabilities for the new model space. Recall that the inclusion probabilities by definition satisfy $\rho_{j|<j}^{(t)} \equiv \sum_{\boldsymbol{\gamma}_{\geq j}} f(\boldsymbol{\gamma}_{\geq j} \mid \boldsymbol{\gamma}_{<j}, \boldsymbol{\rho}^{(t)})\gamma_j$. To update the old inclusion probabilities to the new restricted space $\Gamma_t = \Gamma_{t-1} - \{\boldsymbol{\gamma}^{(t)}\}$, we subtract the mass of $\boldsymbol{\gamma}_{\geq j}^{(t)}$ times $\gamma_j^{(t)}$ in the numerator and re-normalize by dividing by the mass of the new smaller space:

$$\rho_{j|<j}^{(t)} = \frac{\sum_{\boldsymbol{\gamma}_{\geq j}} f(\boldsymbol{\gamma}_{\geq j} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)})\gamma_j - f(\boldsymbol{\gamma}_{\geq j})^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)})\gamma_j^{(t)}}{\sum_{\boldsymbol{\gamma}_{\geq j}} f(\boldsymbol{\gamma}_{\geq j} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)}) - f(\boldsymbol{\gamma}_{\geq j}^{(t)} \mid \boldsymbol{\gamma}_{<j}^{(t)}, \boldsymbol{\rho}^{(t-1)})}.$$

Using the definition of $\boldsymbol{\rho}_{j|<j}^{(t-1)}$, it is straightforward to show that this simplifies to (9), thus

proving that $f(\boldsymbol{\gamma} \mid \boldsymbol{\rho}^{(t+1)})$ gives the probability distribution on the new model space after removing the current sampled model.

The updated probabilities ensure that all previously sampled models receive probability zero and that all models will be sampled in $T = 2^p$ steps. For all other branches that do not include $\boldsymbol{\gamma}^{(t)}$, the $\rho^{(t)}_{j|<j}$ will remain unchanged from the value at the previous iteration $\rho^{(t-1)}_{j|<j}$. The benefit of this representation is that only the $\rho_{j|<j}$'s along the paths of sampled models need to be stored (which is less than $\mathcal{O}(\mathcal{S}_t)$) and only the $\rho^{(t)}_{j|<j}$ on the path of the current sampled model need to be updated at each iteration. By representing the model space as a recursive set of linked lists and effective use of pointers we may efficiently traverse and update probabilities only when required. Pseudo code for the algorithm is given in the Supplemental Materials.

2.4 Sampling without Replacement from Posterior Distributions

In theory any posterior distribution may be decomposed in the form given by (6) allowing us to generate "perfect" samples without replacement from Γ. In practice, the sequence of conditional probabilities are generally unknown unless there is additional structure in the problem, such as posterior independence (Clyde 1999) as in the case of design matrices with orthogonal columns or limited dependence such as a Markov property. Otherwise in the general case, the computational complexity of finding all conditional probabilities for the initial $\boldsymbol{\rho}$ is equivalent to that of enumerating the space.

Nevertheless, we can sample without replacement from a sequence of distributions that are "close" to our target. In the next section, we construct an adaptive algorithm for sampling without replacement from the space of models.

3 BAYESIAN ADAPTIVE SAMPLING

Motivated by the optimality of the median probability model of Barbieri and Berger (2004) and the equivalent centroid model of Carvalho and Lawrence (2008), we sample without replacement using $\{\rho_{j|<j} = \pi_j\}$, where π_j is the marginal posterior inclusion probability of

variable j. Under a posterior model of independence for the inclusion indicators γ_j, the size variables formed with the posterior marginal inclusion probabilities are perfectly correlated with the true posterior model probabilities. While perfect correlation does not hold generally, the model of independence using the current estimates of posterior inclusion probabilities provides a first order approximation to the posterior model probabilities, and, as we prove in Theorem 2 in the Supplemental Materials, is the closest product Bernoulli model to the posterior distribution, where closest is defined in terms of Kullback-Leibler divergence.

If the ensemble of models suggests that a variable is important/unimportant as measured by the marginal inclusion probability, then proposing to include/exclude it based on the marginal inclusion probabilities may be a more efficient way to identify other good models. In the extreme, with a pair of perfectly correlated non-null predictors, the marginal inclusion probabilities will be close to 0.5; sampling with the inclusion probabilities will permit one to make global moves easily and visit the multiple modes.

The marginal posterior inclusion probabilities are unknown prior to sampling (except in the orthogonal case above). We start with an initial estimate and then adaptively update the values using the marginal likelihoods from the sampled models. We first describe the updating scheme, and then suggest possible choices for the initial sampling probabilities.

3.1 Adaptive Updating of the Sampling Inclusion Probabilities

As models are continually sampled, it is appealing to update the sampling inclusion probabilities sequentially with the current estimate of the marginal posterior inclusion probabilities

$$\hat{\pi}_j^{(t)} = \frac{\sum_{\boldsymbol{\gamma} \in \mathcal{S}_t} p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) \gamma_j}{\sum_{\boldsymbol{\gamma} \in \mathcal{S}_t} p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}})} \tag{11}$$

where $\mathcal{S}_t$ is the set of models that have been sampled at time t. Note that when the number of iterations is equal to the population size i.e. $T = 2^p$, these estimates recover the true marginal posterior inclusion probabilities, π_j. This is similar in spirit to a desirable finite sample property called Fisher-consistency (Fisher 1922) where the estimator will recover the population quantities when applied to the entire population.

Adaptive updating does come with a price. If we update $\boldsymbol{\rho}$ with $\hat{\boldsymbol{\pi}}^{(t)}$, we also have to re-normalize the distribution over the binary tree to obtain the new $\{\rho^{(t)}_{j|<j}\}$ that ensure that previously visited models receive zero probability under the new estimates of π_j. Currently, this is the most expensive step in the algorithm as it requires one to retrace the sequence of sampled models and adjust the sampling probabilities over previously sampled branches. A compromise is to sample initially using $\boldsymbol{\rho} = \boldsymbol{\rho}^{(0)}$, then every U iterations update $\boldsymbol{\rho}^{(0)}$ using $\hat{\boldsymbol{\pi}}^{(U)}$ if $\|\hat{\boldsymbol{\pi}}^{(t)} - \hat{\boldsymbol{\pi}}^{(t-U)}\|^2/p > \delta$ for some $\delta > 0$. For large problems, updating periodically when the marginal inclusion probabilities have changed significantly, rather than at every iteration balances computational cost with improved sampling probabilities. Also, care must be taken not to adapt too early, as estimates $\hat{\pi}^{(t)}_j$ may be one or zero, if the corresponding γ_j is always one or zero in the sample $\mathcal{S}_t$. In practice, we bound the sampling inclusion probabilities $\boldsymbol{\rho}$ away from 0 or 1 by constraining $\rho^{(t)}_{j|<j} \in (\epsilon, 1-\epsilon)$, so that all models receive positive sampling probability. We use $\epsilon = .025$ for the bound on the sampling probabilities and set $\delta = \sqrt{\epsilon}$ (based on the average change) for determining whether the update step should be performed.

3.2 Choice of Initial Sampling Inclusion Probabilities

We now discuss some choices for the initial sampling probabilities.

3.2.1 Uniform Probabilities

Setting each $\rho_{j|<j}$ equal to $1/2$ corresponds to equal probability sampling or SRSWOR initially. The estimated marginal inclusion probabilities using (11) after the first U draws are a ratio of Horvitz-Thompson estimates (Horvitz and Thompson 1952) and are approximately unbiased (Thompson 1992, Ch. 6).

3.2.2 P-value Calibration

A simple strategy is to calibrate p-values to Bayes factors and then probabilities using the results of Selke et al. (2001). Consider testing a single precise null hypothesis of the form $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$. If the p-value for the test satisfies $p < 1/e$,

then Selke et al. (2001) show that a lower bound for the Bayes factor for comparing H_0 to H_1 is $-e\,p\log(p)$, leading to an upper bound on $p(H_1 \mid \mathbf{Y}) = (1 - e\,p\log(p))^{-1}$ (under equal prior odds). In the linear model context, there are multiple tests that one could consider, and as a rule p-values from these tests provide lower bounds to Bayes factors for testing $\beta_j = 0$ conditional on the other β_k for $k \neq j$.

For constructing initial sampling probabilities, we fit the full model to the data and obtain p p-values, p_j, for testing $H_{0j} : \beta_j = 0$ versus $H_{1j} : \beta_j \neq 0$ given that the coefficients for the remaining variables are not zero. Under equal prior odds of inclusion, we use the Selke et al. bound to calibrate the p-values to posterior probabilities

$$\rho_{j|<j} = \begin{cases} 1/\{1 - ep_j\log(p_j)\} & \text{if } p_j < 1/e \approx 0.37 \\ \\ 1/2 & \text{otherwise} \end{cases} \tag{12}$$

where as $p_j \to 1/e$ (from below), the upper bound to the probability converges to 1/2. In the case of orthogonal columns in the design matrix, the p-value calibration provides an upper bound to the posterior marginal inclusion probabilities. For non-orthogonal designs we find this p-value calibration still useful. Variables that are important even after adjusting for all other variables will have small p-values, and hence large values for $\rho_{j|<j}$. In the case of highly correlated variables, the associated p-values in the full model may be large, as one may not need to include the variable in the full model given that the other variables are included. For p-values greater than $1/e$, the approximation reverts to uniform sampling initially for those variables. This actually allows the algorithm to explore the multiple modes associated with the inclusion of one of the highly correlated pairs of variables. We denote this the "eplogp" calibration of p-values.

3.2.3 MCMC Estimates

The eplogp calibration is based on p-values from the full model and is an upper bound to a conditional inclusion probability rather than the marginal inclusion probability. An alternative strategy for problems with high correlations among the predictors is to run a

Markov chain to estimates the marginal inclusion probabilities to provide the initial sampling probabilities for BAS. There are two estimators of model and inclusion probabilities that are typically used. The first is based on the ergodic average or Monte Carlo frequencies

$$\hat{p}^{\mathsf{MC}}(\mathcal{M}_\gamma \mid \mathbf{Y}) = \frac{1}{T}\sum_{t=1}^{T} I(\mathcal{M}^{(t)} = \mathcal{M}_\gamma) \tag{13}$$

$$\hat{\pi}_j^{\mathsf{MC}} = \frac{1}{T}\sum_{t=1}^{T} \gamma_j^{(t)} = \sum_{\boldsymbol{\gamma}\in\Gamma} \gamma_j \, \hat{p}^{\mathsf{MC}}(\mathcal{M}_\gamma \mid \mathbf{Y}) = \sum_{\boldsymbol{\gamma}\in\mathcal{U}} \gamma_j \, \hat{p}^{\mathsf{MC}}(\mathcal{M}_\gamma \mid \mathbf{Y}) \tag{14}$$

where $\mathcal{U}$ is the set of unique models that were sampled. As $T \to \infty$, both the estimated model probabilities and inclusion probabilities converge almost surely to $p(\mathcal{M}_\gamma \mid \mathbf{Y})$ and π_j, respectively.

The second approach is based on the estimates of model probabilities normalized over a subset of models (Clyde et al. 1996; George 1999). In (13-14), the probability of any unsampled model is estimated as zero, while the Monte Carlo frequencies for models in $\mathcal{U}$ are noisy versions of the conditional probabilities of models restricted to $\mathcal{U}$. If we replace the Monte Carlo relative frequencies with their expectations conditional on $\boldsymbol{\gamma}$ in $\mathcal{U}$, this leads to

$$\hat{p}^{\mathsf{RM}}(\mathcal{M}_\gamma \mid \mathbf{Y}) \equiv \frac{p(\mathcal{M}_\gamma \mid \mathbf{Y})}{\sum_{\gamma\in\mathcal{U}} p(\mathcal{M}_\gamma \mid Y)} I(\gamma \in \mathcal{U}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\gamma)p(\mathcal{M}_\gamma)}{\sum_{\gamma\in\mathcal{U}} p(\mathbf{Y} \mid \mathcal{M}_\gamma)p(\mathcal{M}_\gamma)} I(\gamma \in \mathcal{U}) \tag{15}$$

$$\hat{\pi}_j^{\mathsf{RM}} = \sum_{\boldsymbol{\gamma}\in\mathcal{U}} \gamma_j \, \hat{p}^{\mathsf{RM}}(\mathcal{M}_\gamma \mid \mathbf{Y}). \tag{16}$$

While biased under repeated sampling, the re-normalized estimates are both Fisher consistent and asymptotically consistent (Clyde and Ghosh 2010).

4 SIMULATED DATA

We compare BAS to SRSWOR and MCMC methods using simulated data with $p = 15$ and $n = 100$ so that the exact posterior model probabilities may be obtained by enumeration of the model space. All columns of the design matrix except the ninth were generated from in-

dependent $\mathsf{N}(0,1)$ random variables and then centered. The ninth column was constructed so that it's correlation with the second column was approximately 0.99. The regression parameters were chosen as $\alpha = 2$, $\boldsymbol{\beta} = (-0.48, 8.72, -1.76, -1.87, 0, 0, 0, 0, 4.00, 0, 0, 0, 0, 0, 0)'$ and $\phi = 1$. For the parameters in each model (1), we use Zellner's g-prior (Zellner 1986; Liang et al. 2008) with $g = n$,

$$p(\alpha, \phi \mid \boldsymbol{\gamma}) \propto 1/\phi, \quad \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \phi \sim \mathrm{N}_{p_{\boldsymbol{\gamma}}} \left(\mathbf{0}, g(\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}/\phi\right) \tag{17}$$

and $p_{\boldsymbol{\gamma}}$ is the rank of $\mathbf{X}_{\boldsymbol{\gamma}}$, which leads to the marginal likelihood of a model proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_{\boldsymbol{\gamma}}) \propto (1+g)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}} \{1 + g(1 - \mathrm{R}_{\boldsymbol{\gamma}}{}^2)\}^{-\frac{(n-1)}{2}} \tag{18}$$

where $\mathrm{R}_{\boldsymbol{\gamma}}{}^2$ is the usual coefficient of determination; with this scaling, the marginal likelihood of the null model is 1.0. To complete the prior specification, we use a uniform prior distribution over the model space, $p(\mathcal{M}_{\boldsymbol{\gamma}}) = 1/2^p$. Under these prior distributions and with the data generated as above, the posterior marginal inclusion probabilities are close to 0.5 for predictor variables two and nine, leading to a bimodal posterior distribution over the model space.

For our simulation study we consider two Metropolis-Hastings algorithms. At iteration t, the MCMC Model Composition or MC^3 algorithm of Madigan and York (1995) and Raftery et al. (1997) uniformly selects a coordinate j at random and then $\boldsymbol{\gamma}^*$ is proposed by setting $\gamma_j^* = 1 - \gamma_j^{(t)}$ with $\gamma_k^* = \gamma_k^{(t)}$ for $k \neq j$ with $\boldsymbol{\gamma}^*$ accepted with probability $\min(1, p(\boldsymbol{\gamma}^* \mid \mathbf{Y})/p(\boldsymbol{\gamma} \mid \mathbf{Y})$. This is equivalent to the antithetic method A in Nott and Green (2004), who show that this has smaller asymptotic variances for ergodic averages than the Gibbs sampler. Because one-at-a time updates may exhibit poor mixing with highly correlated variables, we consider an additional update proposal that randomly selects a variable included in the current model $\boldsymbol{\gamma}^{(t)}$ (if the current model is not the full model or null model) and then swaps it with a randomly selected variable excluded in the current model. For the Random-Swap (RS) algorithm we propose a new state using the MC^3 proposal with probability $\omega(\gamma^{(t)})$ and otherwise use the swap proposal with probability $1 - \omega(\gamma^{(t)})$, where $\omega(\gamma^{(t)}) = 1$ for the full

and null models and is 1/2 otherwise. This is similar to the reversible-jump MCMC method of Denison et al. (1998) (DMS) adopted by Nott and Green (2004) for the variable selection problem. We have not included the Adaptive MCMC (AMCMC) sampler of Nott and Kohn (2005) in this comparison as we found that there was little difference between it and the Metropolized Gibbs sampler (a random scan version of MC^3) used in an earlier version of the manuscript; one explanation is that single block AMCMC sampler uses the past samples to estimate the conditional probability that $\gamma_j = 1$ given the remaining elements $\boldsymbol{\gamma}_{(-j)}$ and $\mathbf{Y}$ (approximating the full conditional in the Gibbs sampler) which lead to similar mixing problems as the other one-at-a-time update schemes. The RS sampler which uses a simple exchange step improves upon the one-at-a-time update schemes allowing one to escape local modes. Nott and Green (2004) show that the DMS algorithm often does well as their more complicated (and harder to automate) Swendsen-Wang algorithms.

4.1 Comparison to SRSWOR

We first compare SRSWOR to BAS using the p-value calibration for determining the initial sampling probabilities $\boldsymbol{\rho}$. The values for $\boldsymbol{\rho}$ are updated using the current estimates of the posterior marginal inclusion probabilities every 500 iterations (at most). We ran both algorithms for the same number of iterations, which varied from 1 to 10 percent of the model space ($2^{15} = 32,678$). Figure 2 shows the box-plots of the total posterior probabilities of unsampled models for BAS and SRSWOR based on 100 repetitions of each algorithm. The posterior probability of unsampled models decreases linearly for SRSWOR, while with BAS it appears to decrease exponentially. Because BAS adaptively updates the sampling probabilities potentially every 500 iterations, its running time is longer than SRSWOR. While SRSWOR can sample 20% of the $2^{15} = 32,768$ models in the same time that BAS can sample 10%, the probability of unsampled models is significantly lower for BAS; around 5% of the mass remains unsampled for BAS compared to approximately 80% for SRSWOR indicating that, even when taking into account running time, BAS is much more effective than SRSWOR at finding high probability models.

4.2 Comparison to MCMC Algorithms

Next, we compare BAS with the MCMC algorithms MC^3 and RS based on running each algorithm for the same number of iterations (10% of the model space or 3,276 iterations). To increase the effective sample size, we also include a thinned version of RS (RS-Thin), based on running p times longer and saving every pth draw. The results based on running each algorithm 100 times with different seeds are summarized in Figure 3. Clearly, BAS outperforms the MCMC algorithms in terms of total posterior mass for the same number of iterations.

Table 1 and Table 2 show estimates of the bias and mean squared error, respectively, for estimating the marginal inclusion probabilities, model probabilities and mean of $\mathbf{Y}$ using the different algorithms in the 100 simulated data sets. For scalar quantities, we report the average bias over the 100 simulations, $\mathsf{bias}(\Delta) = \sum_{i=1}^{100} \left(\mathsf{E}[\hat{\Delta}|\mathbf{Y}]^{(i)} - \mathsf{E}[\Delta \mid \mathbf{Y}] \right) /100$ while for a J dimensional vector, e.g. $\boldsymbol{\mu}$, bias is expressed as $(\sum_j^J (\mathsf{bias}(\Delta_j)^2/J)^{1/2}$. The mean squared error for a scalar Δ is $\mathsf{MSE}(\Delta) = \sum_{i=1}^{100} (\mathsf{E}[\hat{\Delta}|\mathbf{Y}]^{(i)} - \mathsf{E}[\Delta \mid \mathbf{Y}])^2/100$, while for vector quantities we report the average MSE of the components. Among the three MCMC algorithms (MC^3, RS, RS-Thin), the estimates based on ergodic averages are almost uniformly better in terms of bias than estimates based on the re-normalized probabilities; the exception being the estimates with MC^3 for the two inclusion probabilities corresponding to variables 2 and 9, whose correlation was 0.99. The addition of the random swap step effectively eliminates this bias. As expected BAS does exhibit some bias, but this is on the order of 1% for the inclusion probabilities, with a trend of overestimating larger inclusion probabilities, while underestimating smaller inclusion probabilities. Interestingly, the bias in SRSWOR is often of the opposite sign of the other re-normalized estimators, leading to underestimation of inclusion probabilities. For estimating the mean $\boldsymbol{\mu}$ under model averaging BAS has the smallest bias compared to MC^3 and RS (based on the same number of number of iterations).

While the re-normalized estimates under the MCMC algorithms exhibit small variances, the bias term often dominates the MSE, so that there is no clear winner between the re-normalized and MC estimators here. While SRSWOR often has less bias than BAS it has the most variability of any of the methods. On the other hand, BAS has the smallest MSE

compared to any of the MC^3 and RS estimates for the same number of iterations. While RS-Thin using ergodic averages generally has the smallest bias overall, if BAS were run for the equivalent number of proposed model evaluations, BAS would have enumerated the model space and provided exact estimates. Even without enumeration, BAS often has MSEs that are better than RS-Thin.

Quantity	Truth	BAS		MC^3		RS		RS-Thin		SRSWOR
Δ	π_j	eplogp	uniform	MC	RM	MC	RM	MC	RM	
γ_{12}	0.09	-1.23	-1.35	-0.14	-4.21	0.35	-3.80	0.01	-2.54	1.77
γ_{14}	0.10	-1.14	-1.25	-0.23	-4.23	0.05	-3.89	-0.02	-2.44	0.92
γ_{10}	0.11	-1.14	-1.30	-0.10	-4.23	0.11	-4.02	0.00	-2.56	0.93
γ_{8}	0.12	-0.97	-1.11	0.36	-3.94	-0.51	-3.81	0.08	-2.36	1.45
γ_{6}	0.13	-1.05	-1.27	-0.65	-4.64	0.06	-4.24	0.06	-2.57	0.53
γ_{7}	0.14	-1.04	-1.18	-0.13	-4.41	0.08	-4.12	0.06	-2.53	0.04
γ_{13}	0.15	-1.15	-1.24	-0.49	-4.76	0.28	-4.32	0.11	-2.54	0.41
γ_{11}	0.16	-1.13	-1.28	-0.38	-4.59	-0.10	-4.44	-0.05	-2.60	0.82
γ_{15}	0.17	-0.78	-0.92	-0.58	-4.15	-0.19	-3.74	0.09	-2.24	1.56
γ_{5}	0.48	-0.25	-0.38	-0.29	-0.94	0.46	-1.17	-0.12	-0.55	2.32
γ_{9}	0.51	-0.32	-0.26	-1.79	-2.20	-0.22	-1.53	-0.14	-1.04	2.12
γ_{2}	0.54	0.34	0.27	1.73	0.29	0.35	-0.25	0.14	-0.40	-1.45
γ_{1}	0.74	1.19	0.91	-0.23	3.39	0.41	3.69	0.21	2.10	0.17
γ_{3}	0.91	1.56	1.30	-0.40	3.59	-0.14	4.00	0.06	2.35	-0.73
γ_{4}	1.00	0.00	0.00	0.01	0.01	-0.02	0.01	-0.00	0.01	-0.00
$\mathbf{I}(\boldsymbol{\gamma})$	-	3.06	3.29	3.72	22.57	2.95	20.57	1.00	9.23	20.26
$\boldsymbol{\mu}$	-	6.71	5.95	10.62	19.53	10.75	20.35	1.54	11.88	6.63

Table 1: Bias for simulated data calculated from 100 replicates: the values reported in the table are Bias $\times 10^2$ for $\Delta = \gamma_j$, Bias $\times 10^5$ for $\Delta = \mathbf{I}(\boldsymbol{\gamma})$, and Bias $\times 10^3$ for $\Delta = \boldsymbol{\mu}$.

5 U.S. CRIME DATA

We illustrate BAS and the RS algorithm using the US Crime data of Vandaele (1978), which has been considered by Raftery et al. (1997), among others, as a test-bed for evaluating methods for model selection and model averaging. There are 15 predictors and following Raftery et al. (1997), we log transform all continuous variables. For illustration, we use the g-prior with $g = n$ and uniform distribution over the model space as in the simulation study in Section 4.

Quantity	Truth	BAS		MC^3		RS		RS-Thin		SRSWOR
Δ	π_j	eplogp	uniform	MC	RM	MC	RM	MC	RM	
γ_{12}	0.09	1.23	1.35	2.77	4.27	2.14	3.83	0.65	2.54	6.12
γ_{14}	0.10	1.14	1.26	2.92	4.31	2.59	3.95	0.63	2.44	5.26
γ_{10}	0.11	1.15	1.31	3.06	4.31	2.40	4.07	0.68	2.57	5.58
γ_{8}	0.12	0.97	1.12	2.77	4.01	2.23	3.87	0.76	2.37	6.24
γ_{6}	0.13	1.05	1.28	3.12	4.74	2.72	4.31	0.78	2.58	6.28
γ_{7}	0.14	1.05	1.19	3.45	4.52	2.50	4.17	0.78	2.54	5.26
γ_{13}	0.15	1.15	1.24	3.50	4.87	2.44	4.38	1.00	2.55	6.22
γ_{11}	0.16	1.13	1.29	3.64	4.71	3.01	4.52	0.87	2.61	6.86
γ_{15}	0.17	0.78	0.93	3.92	4.27	3.32	3.84	0.79	2.24	7.93
γ_{5}	0.48	0.27	0.40	3.69	1.41	4.35	1.59	1.21	0.60	14.14
γ_{9}	0.51	0.37	0.39	16.70	5.62	6.93	2.08	2.08	1.07	13.35
γ_{2}	0.54	0.39	0.40	16.56	5.25	6.91	1.46	2.15	0.48	13.05
γ_{1}	0.74	1.20	0.92	4.10	3.55	4.51	3.90	1.30	2.11	11.13
γ_{3}	0.91	1.57	1.31	2.96	3.66	3.42	4.10	0.69	2.36	4.48
γ_{4}	1.00	0.00	0.00	0.01	0.01	0.17	0.01	0.03	0.01	0.00
$\mathbf{I}(\boldsymbol{\gamma})$	-	3.16	3.40	33.61	25.35	29.43	22.12	10.27	9.57	156.23
$\boldsymbol{\mu}$	-	6.75	6.01	28.56	21.15	25.53	21.20	5.93	11.94	51.85

Table 2: Square root of the mean square error (RMSE) from the 100 simulated data sets: the values reported in the table are RMSE $\times 10^2$ for $\Delta = \gamma_j$, RMSE $\times 10^5$ for $\Delta = \mathbf{I}(\boldsymbol{\gamma})$, and RMSE $\times 10^3$ for $\Delta = \boldsymbol{\mu}$.

We ran BAS for 2^{15} iterations with the initial ρ_j's determined by the p-value calibration and adaptively updated the sampling inclusion probabilities ρ_j at most every 500 iterations. We ran RS for $15 \cdot 2^{15}$ iterations with every 15th model saved for RS-Thin. Figure 4 contrasts the trace-plots of log marginal likelihoods of sampled models from BAS and RS-Thin. The performance of BAS in terms of finding high probability models is comparable to RS-Thin. While BAS is designed to sample without replacement, the MCMC algorithm revisits models according to their posterior probabilities. Both methods find high probability models relatively quickly, but RS-Thin spends a significant time revisiting them (the points in grey), and never completely explores the space of models, visiting $2,994$ unique models, while BAS sampled all $32,768$ in 1/15th of the number of iterations. Models not sampled by RS-Thin, however, generally receive low posterior probability individually, with a total unsampled mass of 0.037. For this example, BAS has zero variance (and zero bias) for the inclusion probabilities or other quantities and leaves no lingering questions regarding missed pockets

of high probability.

6 PROTEIN ACTIVITY DATA

In this example we compare BAS and the thinned RS sampler (RST) on the protein activity data previously analyzed by Clyde and Parmigiani (1998). Coding the categorical variables by indicator variables and considering all main effects and two-way interactions, and quadratic terms for the continuous variables, the resulting linear model has a total of $p = 88$ potential candidate predictors. The corresponding model space of 2^{88} models poses a challenging model search problem because of high correlations among some variables in the design matrix (maximum correlation is 0.99, with 17 pairs of variables having correlations above 0.95). Motivated by the difficulty of model search using MCMC in this problem Clyde et al. (1996) developed an importance sampling algorithm using orthogonal variables for BMA. Here we return to the problem using the original predictors and the g-prior with $g = n = 96$ to compare BAS to RST.

We ran the RS algorithm for $88 \cdot 2^{20}$ iterations, saving every 88th model for RST. We ran BAS using four different choices for the initial sampling probabilities: uniform probabilities (BAS-uniform), the eplogp calibration (BAS-eplogp), and the RST Monte Carlo estimates of inclusion probabilities (BAS-RST-MC) and the RST re-normalized estimates of inclusion probabilities (BAS-RST-RM). BAS-uniform and BAS-eplogp were run for 2^{20} iterations each, while for the combined BAS-RST-MC and BAS-RST-RM methods we used the first 2^{19} models from RST to estimate marginal inclusion probabilities and then ran BAS for 2^{19} iterations, using only the models from BAS for subsequent inference. For the four variants of BAS we updated the sampling probabilities every 10,000 iterations.

We repeated each procedure ten times, and found that the distribution of log marginals was fairly consistent from run to run (Figure 5) within a method, with BAS-RST-MC exhibiting the greatest variability from run to run. Roughly 65% of the variables have initial sampling inclusion probabilities of 1/2 under BAS-eplogp. Although the maximum log marginal likelihood for RST is higher than that found by BAS-eplogp or BAS-uniform, the middle 50% of the distribution for BAS-eplogp was roughly 5 orders of magnitude higher than RST. The

combination of BAS with the MCMC estimates of initial sampling probabilities consistently finds higher log-marginal likelihoods than either RST, BAS-uniform, or BAS-eplogp, with the median log marginal roughly 7 orders of magnitude higher than RST alone. Note that the models depicted in the 10 boxplots for BAS-RST-MC and BAS-RST-RM do not include any of the models from the corresponding RST sample that were used to estimate the sampling probabilities for BAS. In terms of finding high probability models, the combination of MCMC and BAS is better than either algorithm alone.

It is evident from Figure 5 that there is a difference in the median sampled log marginal likelihoods which may have a substantial impact on the estimates of marginal inclusion probabilities. Figure 6 shows pairwise scatterplots of the estimated posterior inclusion probabilities averaged over the 10 simulations. Despite the fact that the RST-RM and RST-MC estimates are calculated from the same chain, the two estimators are not in perfect agreement, with a number of variables having inclusion probabilities near one or zero for RST-RM while being closer to 0.5 for RST-MC; suggesting that the chains have not converged. The SRSWOR estimates of inclusion probabilities appear to be bounded away from both zero and one (which is consistent with the bias observed in the simulation study). While the RST-MC estimates had the least bias in the simulation study, a surprising feature is that the smaller RST-MC estimates are in close agreement with SRSWOR suggesting that the smaller inclusion probabilities in RS-Thin have not converged. The estimates from BAS-eplogp are in close agreement with BAS-uniform (not depicted), but much less so with the other approaches. While BAS-eplogp and RST-MC agree for variables with large inclusion probabilities, there is poorer agreement for the other variables, where the estimates from BAS are much smaller. There are striking differences in their estimates for variables 11 and 54, with average estimates of $(0.89, 0.14)$ for BAS-eplogp and $(0.50, 0.60)$ for RST-MC. BAS-eplogp shows more variation in these inclusion probabilities across runs with estimates often near one or zero, suggesting multiple modes.

For problems with high correlation, the median probability model (MPM), highest probability model (HPM) and the model closest to BMA can be very different. Using the average of the estimates from RST-MC the MPM has 21 variables, including both variables 11 and

54, and has a log marginal likelihood of around 37.8. The HPM includes 28 variables, with all variables in the MPM except variable 11 and has a log marginal likelihood of 41.8. Despite differences in variables, the predictions under these two models are fairly close, with a correlation of 0.94. While the model averaged predictions vary somewhat depending on the sampling method, the models closest[1] to BMA tend to be more similar to the HPM than the MPM in this case.

Next, we compare the performance of BAS and RST based on leave-one-out cross-validation using model averaging, where the CVRootMSE is defined as

$$\text{CVRootMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2} \tag{19}$$

where $\hat{Y}_{(i)}$ is the predicted value under BMA based on leaving out the i^{th} observation. For each case, we computed the prediction averaging over the top 10,000 models (Table 3). While

Table 3: Square root of the cross validation mean squared error (CVRootMSE) for the protein activity data using SRSWOR, BAS with the four initial inclusion probabilities and the RS-Thin (RST) algorithm using the ergodic average (MC) and re-normalized probabilities (RM).

Algorithm	SRSWOR	BAS uniform	BAS eplogp	BAS-RST MC	BAS-RST RM	RST MC	RST RM
CV-RMSE	0.59	0.51	0.67	0.68	0.71	0.69	0.70

BAS-uniform did not find the highest probability models in the full data set, interestingly it leads to the smallest CVRootMSE of all the cross-validation procedures. Clyde and Parmigiani (1998), reported a CVRootMSE of 0.527 based on the top models found by SSVS, however, their estimate is not based on full cross-validation, as the list of models used in model averaging was obtained by running the MCMC algorithm once for the entire data (due to computational constraints); case-deletion updates were used to provide CV estimates of the model probabilities and predictions conditional on the list of models.

[1]The model that minimizes $|hat\mathbf{Y}_{\mathcal{M}_\gamma} - \hat{Y}_{\mathsf{BMA}}\|^2$ is optimal for selection under squared error loss.

7 DISCUSSION

In this paper we introduced BAS, a novel sampling without replacement algorithm for the Bayesian variable selection problem. Unlike MCMC algorithms, BAS is guaranteed to enumerate the space of models if the number of iterations is equal to the dimension of the model space, 2^p (computational resources permitting). Where enumeration is not feasible, BAS seamlessly transitions to a stochastic sampling algorithm. BAS with various choices for the initial inclusion probabilities is implemented in C in the `R` package `BAS`, available from CRAN or `http://www.stat.duke.edu/~clyde/BAS`. While we have used the g-prior with $g = n$ for the examples in this paper, the BAS package also includes mixtures of g-priors such as the well-known Zellner and Siow (1980) Cauchy prior distribution on $\boldsymbol{\beta}_g$ or the hyper-g distribution of Liang et al. (2008), as well as alternative prior distributions on the models. We are working currently on extensions to include generalized linear models using Laplace approximations for marginal likelihoods.

In the simulation study where we could enumerate to confirm results, using BAS to sample a fraction of the model space resulted in improved MSE for various quantities of interest over standard MCMC algorithms, suggesting that in modest problems that preclude enumeration, BAS has a competitive advantage over the MCMC algorithms considered. In the higher dimensional example where the random swap MCMC algorithm almost sampled without replacement, we found that estimates of inclusion probabilities varied greatly across the different methods. While we can be confident that the variables that had large inclusion probabilities across all of the methods are likely important predictors, the high correlations among some of the variables may dilute inclusion probabilities making them less useful as a posterior summary of variable importance, i.e. an inclusion probability may be small because the variable is unimportant or because there are several other correlated variables that "dilute" the posterior mass among the competing variables (George 1999). Although estimates of inclusion probabilities with BAS are likely biased in this case, BAS was able to identify a substantial number of models with high marginal likelihoods, and BMA using BAS had better out of sample performance than the MCMC only methods considered here.

The protein example suggest that the use of MCMC to construct sampling probabili-

ties for BAS leads to some improvements for exploring higher dimensional problems in the presence of strong correlations. Incorporating local proposals, as in MCMC, that explore a neighborhood around a previously sampled model (HPM, MPM, or random selection) may prove to be beneficial in higher dimensional problems, although it is more difficult to construct while sampling without replacement. For highly correlated design matrices, sampling from the marginal inclusion probabilities may be inefficient in higher dimensions. Theorem 1 suggests that more efficient sampling designs may be constructed by incorporating dependence in the initial $\boldsymbol{\rho}$. The sequence of conditional inclusion probabilities could be estimated from an initial MCMC sample in the spirit of the adaptive MCMC algorithm of Nott and Kohn (2005) and adapted as sampling progresses, leading to a global approximation to the joint posterior distribution. Further comparisons with state-of-the art algorithms (Nott and Green 2004; Liang and Wong 2000; Bottolo and Richardson 2008) that incorporate more complex global moves will be useful for high dimensional problems.

While estimates from BAS are Fisher consistent, estimators based on re-normalized probabilities, such as (11), used in BAS or with MCMC that ignore the sampling mechanism will be biased, with the bias going to zero as more models are sampled. In principle, the Horvitz-Thompson estimator may be used to construct unbiased estimates of the numerator and denominator of (3) by weighting each model's marginal likelihood inversely to the probability that they are included in the sample (similar to importance sampling reweighting in sampling with replacement). Unfortunately, with MCMC or PPS sampling algorithms the probability that a model is included in the sample is generally unavailable. Construction of alternative estimators to adjust for sampling bias and studying their theoretical properties in higher dimensional problems is another important direction for future work.

Finally, because of the binary tree structure in BAS, it is straightforward to implement the algorithm in parallel. With 2^k available processors, one may enumerate the model indicators for the first k variables, and then run BAS conditioning on these variables by setting their inclusion probabilities to γ_j for $j = 1, \ldots k$ in each of the 2^k sub-branches. Using the `snow` package in `R` this may be accomplished with the existing code. We expect significant improvements in speed are also possible with an implementation using modern

graphics processing units.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation grants DMS-9733013 and DMS-0422400 and AST-0507481 and National Institutes of Health grant 1R01-HL-090559. The second author was partially supported by the National Institutes of Health grant NIH/NIEHS 5T32ES007018.

SUPPLEMENTAL MATERIALS

`R-package:` R-package `BAS` to perform the adaptive sampling methods described in the article. The package includes the protein construct data. (GNU zipped tar file, BAS_0.90.tar.gz)

`Data and Code:` Data (simulated and real), `C` and `R` code for MCMC sampling algorithms, post-processing samples and creating figures described in the article. (zip file containing the data, code and a read-me file (readme.pdf), code-rev2.zip)

`Proofs and Pseudo code:` Proofs of theorems in the paper and Pseudo code for the BAS algorithm. (proofs+pseudocode.pdf)

References

Barbieri, M. and Berger, J. (2004), "Optimal predictive model selection," *Annals of Statistics*, 32, 870–897.

Bottolo, L. and Richardson, S. (2008), "Evolutionary Stochastic Search," Technical report, Imperial College, London.

Carvalho, L. E. and Lawrence, C. E. (2008), "Centroid estimation in discrete high-dimensional spaces with applications in biology," *Proceedings of the National Academy of Sciences*, 105, 3209–3214.

Clyde, M. (1999), "Bayesian model averaging and model search strategies (with discussion)," in *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, pp. 157–185.

Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction via orthogonalized model mixing," *Journal of the American Statistical Association*, 91, 1197–1208.

Clyde, M. and George, E. I. (2004), "Model uncertainty," *Statistical Science*, 19, 81–94.

Clyde, M. and Ghosh, J. (2010), "A note on the bias in estimating posterior probabilities in variable selection," Discussion Paper 2010-11, Duke University Department of Statistical Science.

Clyde, M. A. and Parmigiani, G. (1998), "Protein construct storage: Bayesian variable selection and prediction with mixtures," *Journal of Biopharmaceutical Statistics*, 8, 431–443.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society, Series B*, 60, 333–350.

Fisher, R. A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309–368.

Furnival, G. M. and Wilson, J., Robert W. (1974), "Regression by leaps and bounds," *Technometrics*, 16, 499–511.

George, E. (1999), "Discussion of "Model Averaging and Model Search Strategies" by M. Clyde," in *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting.*

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–374.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian model averaging: a tutorial (with discussion)," *Statistical Science*, 14, 382–401, corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Horvitz, D. and Thompson, D. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008), "Mixtures of g-priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103, 410–423.

Liang, F. and Wong, W. H. (2000), "Evolutionary Monte Carlo: Applications to C_p Model Sampling and Change Point Problem," *Statistica Sinica*, 10, 317–342.

Madigan, D. and York, J. (1995), "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.

Nott, D. J. and Green, P. J. (2004), "Bayesian Variable Selection and the Swendsen-Wang Algorithm," *Journal of Computational and Graphical Statistics*, 13, 141–157.

Nott, D. J. and Kohn, R. (2005), "Adaptive sampling for Bayesian variable selection," *Biometrika*, 92, 747–763.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, 92, 179–191.

Selke, T., Bayarri, M., and Berger, J. (2001), "Calibration of P-values for testing precise null hypotheses," *The American Statistician*, 55, 62–71.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–343.

Thompson, S. K. (1992), *Sampling*, Wiley Interscience.

Vandaele, W. (1978), "Participation in illegitimate activities - Ehrlich revisited," in *Deterrence and Incapacitation*, eds. A. Blumstein, J. Cohen, and D. Nagin, Washington D.C.: National Academy of Sciences Press, pp. 270–335.

Wilson, M., Iversen, E., Clyde, M., Schmidler, S., and Schildkraut, J. (2010), "Bayesian Model Search and Multilevel Inference for SNP Association Studies," *Annals of Applied Statistics*, 4, to appear.

Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North-Holland/Elsevier, pp. 233–243.

Zellner, A. and Siow, A. (1980), "Posterior odds ratios for selected regression hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pp. 585–603.

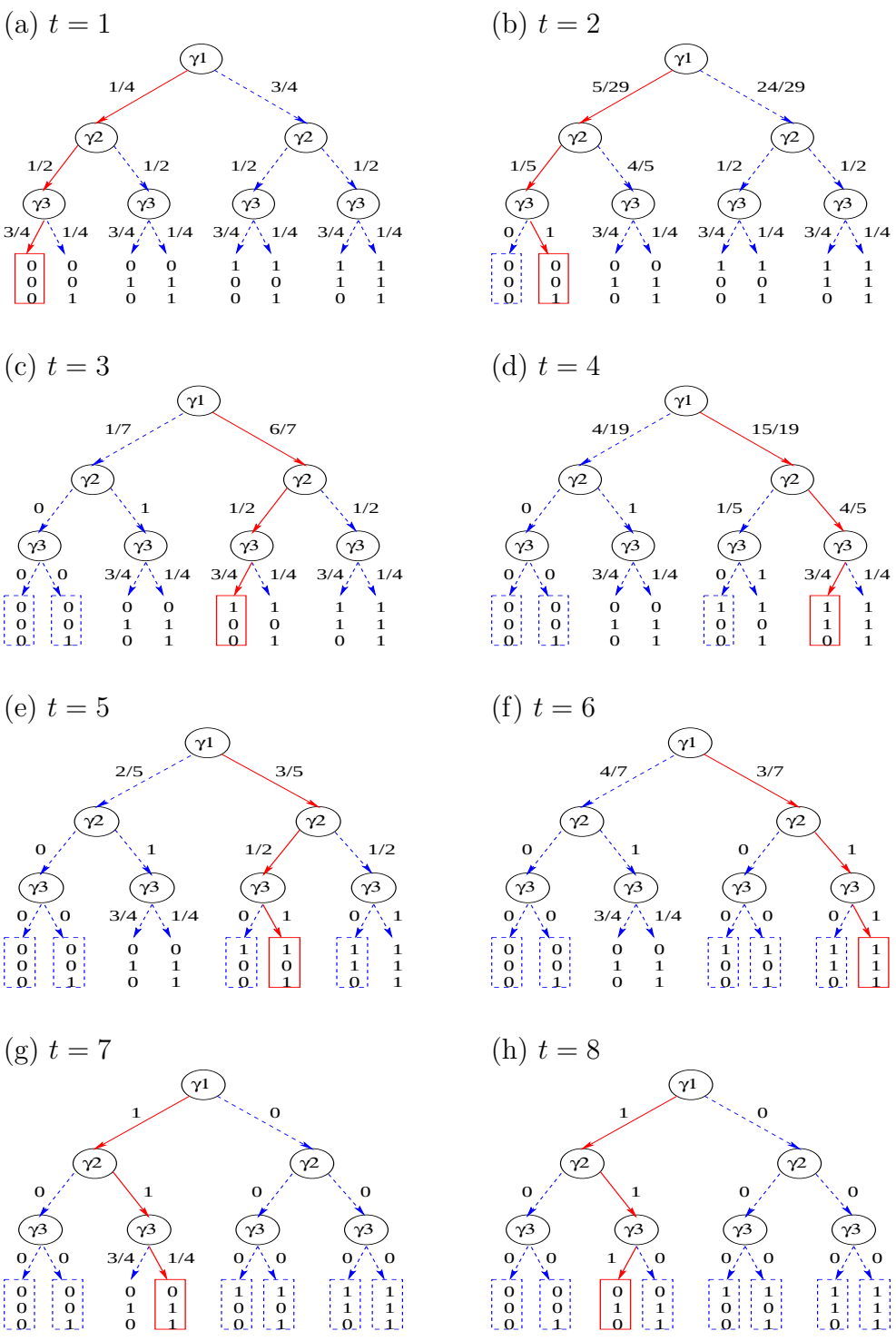

Figure 1: An Illustration of the BAS algorithm for $p = 3$ with inclusion probabilities plotted along the branches of the tree; here solid lines and rectangles represent the model sampled at the t^{th} draw, and dashed rectangles represent models sampled at previous draws.

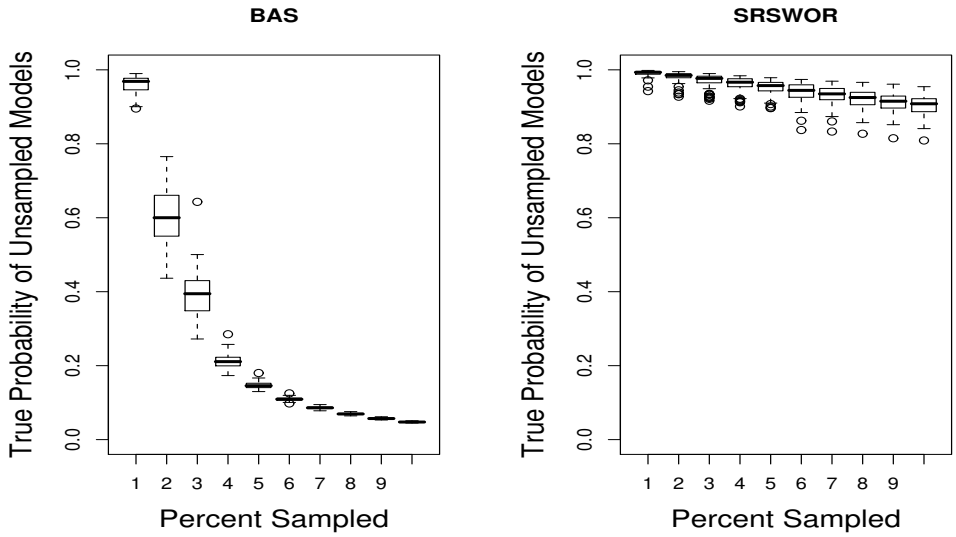

Figure 2: Comparison of BAS and SRSWOR based on the same number of iterations for the simulated data. Box-plots are based on 100 replicates of each algorithm.

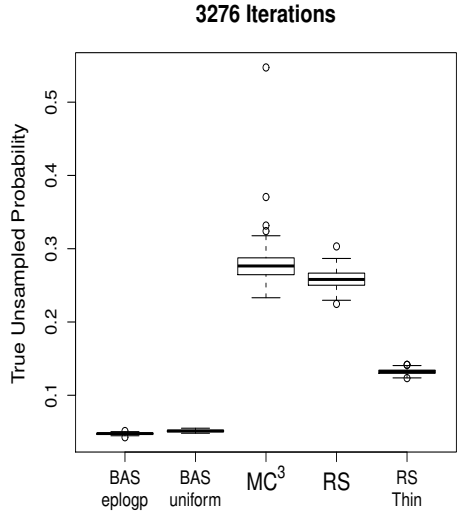

Figure 3: Box-plots of unsampled mass for BAS MC^3, RS and RS-Thin based on 100 replications of each algorithm. The number of iterations for BAS, $MC^3$and RS were equal to 10% the dimension of the model space (2^{15}), while RS-Thin was run 15 times longer, with every 15th model saved.

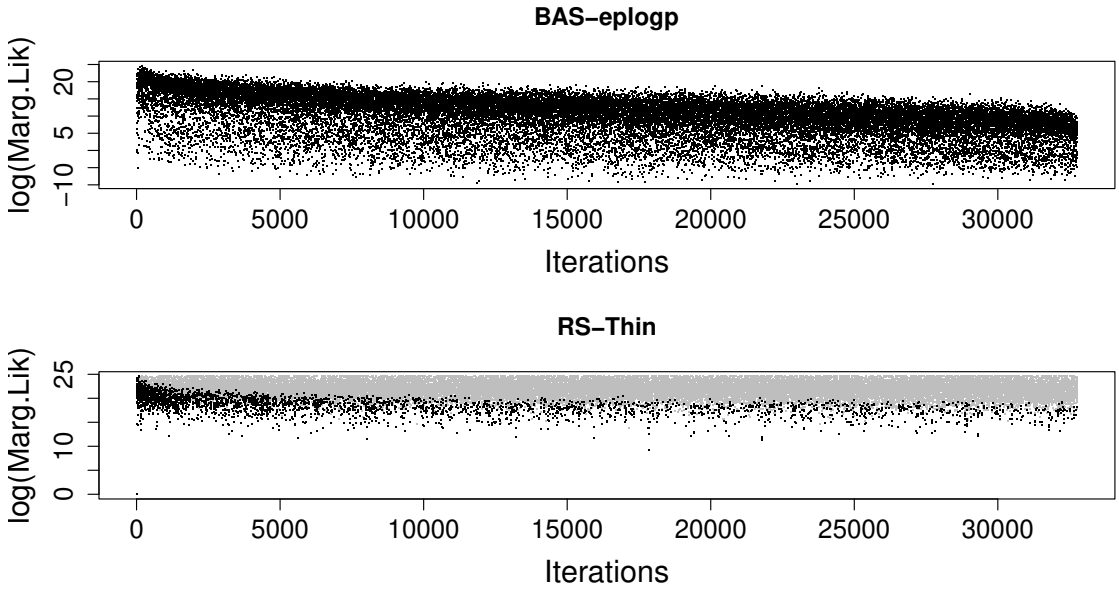

Figure 4: Trace plots in the U.S. Crime data for BAS (2^{15} iterations) and RS-Thin ($15 \cdot 2^{15}$ iterations, saving every 15th model). Black points correspond to the first visit of a model while grey points (in RS-Thin) correspond to models that have been revisited by the Markov chain.

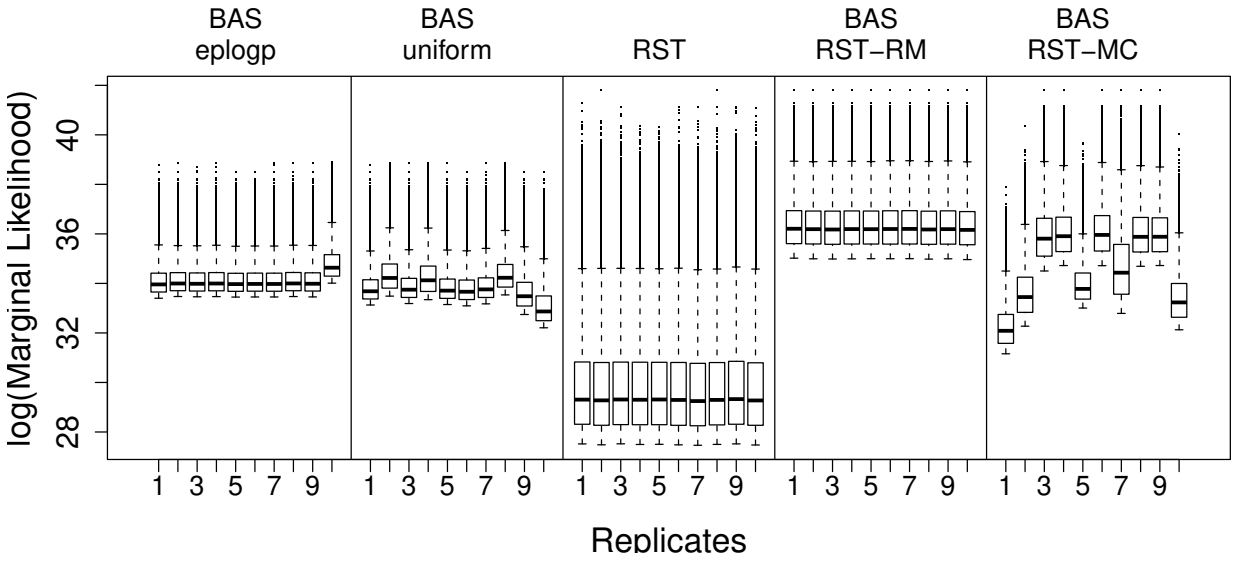

Figure 5: Comparison of the log-marginal likelihood in the protein data of the top 100,000 unique models visited by BAS-eplogp, BAS-uniform, thinned version of Random Swap (RST), BAS with Monte Carlo estimates of inclusion probabilities from the RST samples (BAS-RST-MC), and BAS with re-normalized estimates of inclusion probabilities (BAS-RST-RM) from the RST samples.

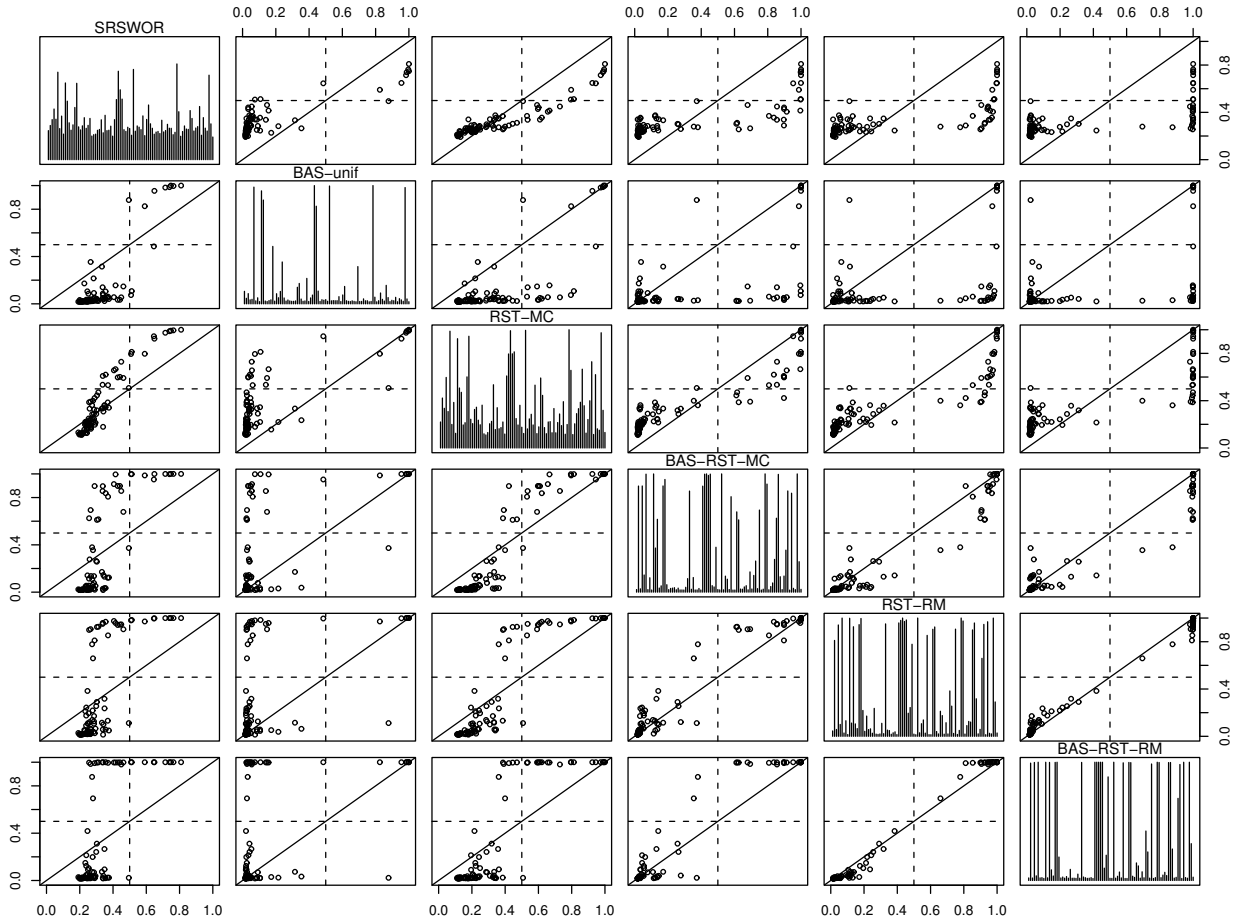

Figure 6: Comparison of estimates of marginal posterior inclusion probabilities for the protein data using simple random sampling without replacement (SRSWOR), BAS with uniform sampling probabilities, (BAS-unif), Monte Carlo estimates from the thinned version of Random Swap (RST-MC), BAS with RST-MC initial sampling probabilities (BAS-RST-MC), re-normalized estimates of inclusion probabilities from the thinned Random Swap (RST-RM) and BAS with initial RST-RM sampling inclusion probabilities (BAS-RST-RM). The points are based on the average of the 10 replicates from each run.