An MCMC model search algorithm for regression problems

Athanassios Petralias and Petros Dellaportas *

Department of Statistics, Athens University of Economics and Business, Greece

Abstract

We improve upon the Carlin and Chib MCMC algorithm that searches in model and parameter space. Our proposed algorithm attempts non-uniformly chosen 'local' moves in model space and avoids some pitfalls of other existing algorithms. In a series of examples with linear and logistic regression we report evidence that our proposed algorithm performs better than existing algorithms.

Keywords: gene selection; population sampling; reversible jump; tracking portfolio

1 Introduction

In Bayesian statistics, model determination is based on the posterior model probability $f(m|y)$ of model m conditional on data y. If m is one of M competing models specifying a distribution $f(y|\beta_m, m)$ conditional on an unknown parameter vector β_m, then this posterior probability is calculated, using Bayes theorem, by

$$f(m|y) \;=\; \frac{f(y|m)f(m)}{\sum\limits_{m\in M} f(y|m)f(m)}, \qquad m \in M \tag{1}$$

where $f(m)$ is the prior probability of model m, $f(\beta_m|m)$ is the prior distribution of the model parameters for model m, and $f(y|m)$ is the marginal likelihood calculated using $f(y|m) = \int f(y|\beta_m, m)f(\beta_m|m)d\beta_m$.

*Correspondence to: Petros Dellaportas, Department of Statistics, Athens University of Economics and Business, Patission 76, Athens 10434, Greece. Email: petros@aueb.gr

We focus on regression problems with p covariates in which each model corresponds to a set of variables so that $|M| = 2^p$ and the problem is equivalent to a variable selection problem. Bayesian model determination is important since calculation of a set of models with highest posterior model probabilities (1) may be extremely useful for both exploratory and predictive purposes. When the number of models $|M|$ is very large, calculation of all posterior model probabilities is computationally prohibitive so MCMC algorithms that search efficiently in model space are necessary. In this paper we propose a new algorithm to perform such a task.

Our suggested algorithm can be viewed as a combination of the Carlin and Chib algorithm (Carlin and Chib, 1995), the 'Metropolised' Carlin and Chib algorithm (Godsill, 2001; Dellaportas et al., 2002) and the Shotgun stochastic search (Hans et al., 2007) algorithm. We call it a 'Subspace Carlin and Chib' (SCC) algorithm and its main idea is based on the following arguments. In the Carlin and Chib algorithm, a Gibbs sampler samples parameters β_k, $k \in M$, and models m from their full conditional densities. Due to the nature of the parameter space the specification of these densities requires the use of linking or pseudoprior densities. Since each iteration requires the generation of all variables from $|M|$ models, this algorithm is infeasible when p is large. A way to overcome this problem is to use the 'Metropolised' Carlin and Chib. There, the Gibbs sampling step in the model space is replaced by a Metropolis step in which a new model is proposed uniformly among all 'local' models that are close to each other, in some sense, in the model space. In variable selection problems this locality is usually specified by models that differ by one or two variables, so they can be approached by adding, deleting or replacing a variable from the current model. In each iteration, the usual Metropolis-Hastings ratio requires the generation of only two parameters and therefore the algorithm becomes very efficient. The price paid is that the nice Gibbs-based global move in model space is replaced by a local move that may be result to an extremely slow mixing Markov chain.

In our proposed SCC algorithm, we first generate parameters β_k, from the current model ($k = m$) posterior density and from all pseudoprior densities of the local models ($k \neq m$). Then, we construct a non-uniform discrete proposal density of all local candidate models based on the generated values of the parameters. Thus, we use a Carlin and Chib algorithm in which model space moves are only allowed in a subspace of M. Compared with similar Markov chains that attempt such 'local' moves, the resulting non-uniformity of the proposal density in model space allows better mixing and can be extremely important when p is large.

In a series of Monte Carlo experiments and applications to real data examples, we tested our algorithm against other competing algorithms by comparing their mean recurrence time. Since the computational cost of iterations differs in different methods, we also used computing time and posterior density evaluations as comparison yardsticks. Moreover, we tested how well our algorithm estimates the true posterior model probabilities when compared

with other algorithms; all algorithms executed the same number of posterior density evaluations. Our Monte Carlo experiments are very encouraging, indicating that in such problems SCC algorithm may be considered as a strong candidate for Bayesian model determination problems.

The paper is organized as follows. In Section 2 we present the Algorithm. Sections 3 and 4 present simulated and real data examples based on linear regressions, whereas Section 5 presents a logistic regression real data example. Finally, Section 6 concludes.

2 The subspace Carlin and Chib algorithm

Assume that we are dealing with a regression problem with p variables and the current state of an MCMC algorithm that samples from the posterior distribution $f(\beta_m, m|y)$ is (m, β_m). Denote prior densities $f(\beta_m, m) = f(\beta_m|m)f(m)$ and pseudopriors $f(\beta_k \mid m \neq k)$. We call the set of all 'local' models to m as neighborhood of m, denoted by S_m, and we split this set into three sub-neighborhoods $S_m = \{S_m^-, S_m^0, S_m^+\}$ representing the sets with variables obtained by deleting, replacing or adding a variable to model m. With a slight abuse of terminology we will order the corresponding sub-neighborhoods of m' as $S_{m'} = \left\{S_{m'}^+, S_{m'}^0, S_{m'}^-\right\}$ so that there is a connectivity between corresponding sub-neighborhoods in the sense that if m' belongs in a sub-neighborhood of S_m then m belongs in the corresponding sub-neighborhood of $S_{m'}$. Our MCMC algorithm proposes a new model $m' \in S_m$ by first choosing one of the three sub-neighborhoods with corresponding probabilities from the set $Q_m = \{q_m^-, q_m^0, q_m^+\}$. Define the corresponding set of probabilities $Q_{m'} = \left\{q_{m'}^+, q_{m'}^0, q_{m'}^-\right\}$. Then, the basic SCC algorithm is described as follows.

Algorithm 1

1. Generate parameters β_k, $k = 1, \ldots, p$, from the full conditional densities
$$f(\beta_k \mid \beta_{l\neq k}, m, y) \propto \begin{cases} f(y \mid \beta_m, m)f(\beta_m, m) & k = m \\ f(\beta_k \mid m \neq k) & k \neq m \end{cases} \tag{2}$$
2. Choose a sub-neighborhood $s_m \in S_m$ with probability $q_m \in Q_m$
3. Propose a new model $m' \in s_m$ with probability
$$j(m, m') = \frac{A_{m'}}{\sum_{\ell \in s_m} A_\ell}, \tag{3}$$
where
$$A_m = f(y \mid \beta_m, m)f(\beta_m \mid m) \prod_{k \neq m} \left[f(\beta_k \mid m \neq k)\right] f(m). \tag{4}$$

4. Accept the proposed model m' with probability

$$\alpha = min\left(1, \frac{A_{m'} q_{m'} j(m', m)}{A_m q_m j(m, m')}\right) = min\left(1, \frac{q_{m'} \sum_{\ell \in s_m} A_\ell}{q_m \sum_{\ell \in s_{m'}} A_\ell}\right). \tag{5}$$

Assume that model m has $0 \leq \kappa \leq p$ variables. In the addition and deletion moves the summations in (3) and (5) require p evaluations of A_ℓ, whereas for the replacement move $(0 < \kappa < p)$ we need $\kappa(p-\kappa)+(\kappa-1)(p-\kappa-1)$ evaluations. Therefore, the replacement move is likely to be inefficient as, for example, when $\kappa = p/2$, the number of required evaluations of A_ℓ is $(p/2)^2 + (p/2-1)^2$ which increases quadratically with p. Therefore, we suggest replacing the Algorithm 1 with the following Algorithm 2 which proposes models uniformly when a replacement move is chosen.

Algorithm 2

1. Choose a sub-neighborhood $s_m \in S_m$ with probability $q_m \in Q_m$
2. If an addition or deletion move is chosen ($s_m \in \{S_m^-, S_m^+\}$) then
   (a) Generate parameters β_k, $k = 1, \ldots, p$, from the full conditional densities

$$f(\beta_k \mid \beta_{l \neq k}, m, y) \propto \begin{cases} f(y \mid \beta_m, m) f(\beta_m, m) & k = m \\ f(\beta_k \mid m \neq k) & k \neq m \end{cases}$$

   (b) Propose a new model $m' \in s_m$ with probability,

$$j(m, m') = \frac{A_{m'}}{\sum_{\ell \in s_m} A_\ell},$$

   where

$$A_m = f(y \mid \beta_m, m) f(\beta_m \mid m) \prod_{k \neq m} \left[f(\beta_k \mid m \neq k)\right] f(m).$$

   (c) Accept the proposed model m' with probability,

$$\alpha = min\left(1, \frac{A_{m'} q_{m'} j(m', m)}{A_m q_m j(m, m')}\right) = min\left(1, \frac{q_{m'} \sum_{\ell \in s_m} A_\ell}{q_m \sum_{\ell \in s_{m'}} A_\ell}\right).$$

3. If a replacement move is chosen ($s_m = S_m^o$) then

   (a) Propose a new model $m' \in S_m^o$ with probability,

   $$j(m, m') = |S_m^o|^{-1}$$

   (b) Accept the proposed model m' with probability,

   $$\alpha = min\left(1, \frac{q_{m'} A_{m'}}{q_m A_m}\right).$$

Algorithms 1 and 2 have a similar philosophy with the MCMC version of the Shotgun stochastic search Algorithm of Hans et al. (2007) but instead of attempting to move directly to all models in S_m, they propose more local moves. Thus, they are more flexible in avoiding 'sticky patches' in model space; see the discussion in, for example, Dellaportas and Roberts (2003). Indeed, Hans et al. (2007) point out that the chain may fail to move if it starts at a region of zero posterior probability mass. In more simulation studies this has been also noted by Petralias (2010).

3 Simulation studies

We first perform a series of simulated data examples based on normal linear regression of the form $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, where y and ϵ are vectors of size $n \times 1$, $\beta = (\beta_1, \ldots, \beta_p)'$ is a set of regressors of size $p \times 1$ and the $n \times p$ design matrix X has columns denoted by X_i, $i = 1, \ldots, p$.

We consider three simulated datasets with $p = 11, 16$ and 21 regressors and we simulate $n = 100$ data points. All error terms are generated from a standard normal density. For the first dataset we set X_1 as a vector of ones and β_i, $i = 7, \ldots, 11$ as vector of zeroes. Then we simulate the elements of design matrix and of β_i, $i = 1, \ldots, 6$, from $N(0, 1)$ density. In the second dataset we set X_1 as a vector of ones and β_i, $i = 14, \ldots, 16$ as vector of zeroes. We then simulate $X_i \sim N(0, 1)$, $i = 2, \ldots, 13$ and impose correlations by setting $X_{14} = \sum_{i=2}^{4} (X_i) + u$, $X_{15} = \sum_{i=5}^{7} (X_i) + u$, $X_{16} = \sum_{i=2}^{7} (X_i) + u$, with u being vectors simulated from a $N(0, 1)$ density. Finally, we generate $\beta_i \sim N(0, 1), i = 1, ..., 13$. The third dataset serves as a replication of the first with more regressors; we generate $\beta_i \sim N(0, 1)$, $i = 1, \ldots, 11$ and set $\beta_i = 0$, $i = 12, \ldots, 21$.

For the priors over the model space, we adopt a Beta-Binomial prior on the number of variables in each model suggested by Kohn et al. (2001), with parameters chosen so that the prior mean and variance of the number of model parameters are one. These prior

specifications are very flexible for a series of reasons; see, for example, Ley and Steel (2009). As in Dellaportas et al. (2002), we use $\beta_i \sim N(0, 10^2)$ for all parameters $i = 1, \ldots, p$ and $\sigma^2 \sim IG(10^{-4}, 10^{-4})$.

We report results only for Algorithm 2 which we have found to be more efficient than Algorithm 1. The pseudopriors $f(\beta_k \mid m \neq k)$ are taken to be independent normal densities with mean and variance obtained through a pilot Gibbs run of 1000 iterations on the saturated model after a burn-in of 100 iterations. We compare it with the following competing algorithms.

- A standard reversible jump algorithm (Green, 1995) with proposal densities taken as our pseudopriors and models in S_m proposed uniformly.
- A population reversible jump algorithm with three parallel chains, two of them being tempered, following guidelines of Jasra et al. (2007b) and Bottolo and Richardson (2010); in particular, we randomly propose a mutation, an exchange or a crossover move. The temperature ladder for the parallel chains is set so as the exchange move is accepted about half of the time as suggested by Liu (2001). More specifically we set the inverse temperature parameter for the chain i, $i = 1, 2s, 3$, with target distribution $\pi_i \propto \pi^{\zeta_i}$, equal to $\zeta_i = z^{i-1}$, with $z = 0.6$ for the dataset with 11 regressors, $z = 0.72$ for the dataset with 16 regressors and $z = 0.74$ for the dataset with 21 regressors. Thus, tempered chains are expected to increase the mixing of the algorithm in the model space.
- An MC^3 algorithm (Madigan and York, 1995) which utilizes the same move types as the reversible jump algorithm but samples directly form the model space after marginalising out the parameters in each model.
- A 'Metropolised' Carlin and Chib algorithm as described in Dellaportas et al. (2002) with a a multivariate proposal (pseudoprior) density

$$N\left(\left(X_{m'}^T X_{m'}\right)^{-1} X_{m'}^T y, \left(X_{m'}^T X_{m'}\right)^{-1} \hat{\sigma}^2\right),$$

  where $\hat{\sigma}^2$ is the current value of the residual variance and X_m denotes the design matrix of model m.
- A Gibbs variable selection algorithm as described in Dellaportas et al. (2002).

All algorithms were independently run 100 times from randomly chosen initial points and the following diagnostics were used for comparison. First, the mean recurrence time, see for example Grimmett and Stirzaker (2004), of visiting the highest posterior probability

Algorithm	Sample taken	Time (seconds)	Acceptance rate			Mean recurrence time			MSE
			Add	Delete	Replace	Iter.	Sec.	Eval.	
First dataset with 11 regressors									
SCC	11,550	16.61	0.3235	0.3240	0.0043	9.93	0.0143	85.94	0.0100
RJ	50,000	19.94	0.0607	0.0608	0.0038	57.22	0.0228	114.45	0.0119
Pop. RJ	18,942	19.91	0.0610	0.0603	0.0041	104.32	0.0393	197.48	0.0157
MC^3	50,000	17.51	0.0950	0.0950	0.0041	35.38	0.0124	70.76	0.0096
MCC	33,334	30.56	0.0703	0.0703	0.0043	49.26	0.0452	147.77	0.0138
GVS	8,334	21.03		0.4627		4.89	0.0123	58.64	0.0065
Second dataset with 16 regressors									
SCC	8,342	18.81	0.3632	0.3637	0.0096	28.26	0.0637	338.76	0.0164
RJ	50,000	23.62	0.0639	0.0640	0.0089	171.92	0.0812	343.85	0.0188
Pop. RJ	18,689	22.03	0.0637	0.0631	0.0090	201.02	0.0840	380.34	0.0255
MC^3	50,000	20.88	0.0938	0.0938	0.0121	140.62	0.0587	281.25	0.0152
MCC	33,334	36.52	0.0787	0.0784	0.0120	148.53	0.1628	445.58	0.0190
GVS	5,883	23.41		0.5112		12.34	0.0491	209.76	0.0106
Third dataset with 21 regressors									
SCC	6,526	20.61	0.1913	0.1920	0.0035	20.44	0.0646	313.28	0.0161
RJ	50,000	25.52	0.0204	0.0204	0.0015	218.73	0.1116	437.45	0.0205
Pop. RJ	19,082	25.14	0.0198	0.0200	0.0018	224.24	0.1076	427.62	0.0215
MC^3	50,000	22.64	0.0285	0.0286	0.0022	176.29	0.0798	352.57	0.0140
MCC	33,334	40.89	0.0258	0.0259	0.0024	181.89	0.2231	545.67	0.0188
GVS	4,546	27.06		0.3009		8.74	0.0520	192.30	0.0095

Table 1: Simulation results: SCC: Subsample Carlin and Chib; RJ: reversible jump; Pop. RJ: population reversible jump (acceptance rates are displayed for the first untempered chain); MCC: 'Metropolised' Carlin and Chib; GVS: Gibbs variable selection (acceptance rates are reported for the global model selection move)

model is calculated. This diagnostic may be considered as the average iterations needed to revisit the best model. For full generality we present it in terms of iterations, CPU time and posterior density evaluations. Moreover, since in linear models the true posterior model probabilities are analytically available, see for example O'Hagan and Forster (2004), we estimated the mean standard error (MSE) of the estimated probability of every model with posterior model probability above 0.001 and we report the weighted average of all mean squared errors. Since each iteration required different computer power, each algorithm ran for the same number $(100,000)$ of posterior density evaluations.

The simulation results for all datasets and algorithms considered are presented in Table 1. The reported numbers have been obtained as sample averages over 100 independent resplications. Except for MC^3 which naturally beats Metropolis-based algorithms in terms of mean squared error, the SCC Algorithm performs really well, producing the smallest mean standard error and mean recurrence time against all other competitors except Gibbs variable selection. Note that Gibbs variable selection can only be used when parameter prior independence is assumed. Furthermore, the SCC turns to be the algorithm that completes faster (in terms of seconds) the specified number of posterior evaluations.

4 An SP500 tracking portfolio

We use a dataset of the SP500 index obtained from Bloomberg, for the period 3/1/2007 up to 31/12/2007. The problem addressed here is to identify a relative small collection of stocks that tracks the SP500 index. A portfolio of such stocks is usually called a Financial Index Tracking Portfolio; see also George and McCulloch (1997). We employ a normal linear regression model with response the index returns and design matrix of size 250×435 that represents returns of $p = 435$ stocks for $n = 435$ days; stocks with missing values during that period have been excluded from the analysis. Inspection of the sample correlation matrix reveals that there are about 80 pairs of stock returns with absolute correlation greater than 0.8.

For the coefficient parameters of the linear regression we employ a g-prior with $g = \max(n, p^2)$, see Fernandez et al. (2001), and for the residual variance an Inverse Gamma with hyperparameters 2×10^{-3} and 2×10^{-6}. The prior density on the number of variables included in each model is taken to be Beta-Binomial with parameters 1.2538 and 108.08, chosen so that the prior mean and standard deviation of the model size are 5; see Ley and Steel (2009). Furthermore, we impose the restriction that the maximum number of variables included in the model are less than $n = 250$.

Since we deal with a linear regression model we integrate out all parameters and choose to use population based algorithms that sample in the model space. We have found that convergence without the adoption of tempered chains is extremely slow. We compare a population reversible jump against the Algorithm 2 by using 10 parallel chains. This number turned out to be optimal by following the guidelines of Jasra et al. (2007b).

We compared the algorithms by first running the population reversible jump sampler for three million iterations and then running the population SCC algorithm for as long as the same posterior density evaluations are performed. Although the exact posterior model probabilities are not available analytically since they require an enormous number of marginal density calculations, posterior odds of the most probable models detected by both algorithms can be routinely calculated so weighted average mean standard errors of the estimated probability odds over the 20 best models can be reported. Note that both algorithms had the 20 best models in the list of the 100 models with the highest posterior probability. Some diagnostics are presented in Table 2 whereas results associated with the posterior probability odds are presented in Table 3. SCC turns out to be better in mean recurrence times and estimated posterior odds standard errors.

Figure 1 depicts the ergodic probability for the highest posterior probability model and the estimated posterior odds of the three next models against the first. It is evident that SCC performs better.

The twenty highest posterior probability models have five to seven variables. The highest

	Population reversible jump	Population SCC
Sample taken	3,000,000	232,610
Overall time (seconds)	19,538	6,177
Overall evaluations	29,821,350	29,821,563
Average acceptance rate of addition moves	0.0621	0.3683
Average acceptance rate of deletion moves	0.0623	0.3692
Average acceptance rate of replacement moves	0.0168	0.0174
Average acceptance rate of exchange moves	0.4684	0.4711
Average acceptance rate of crossover moves	0.0275	0.0264
Mean recurrence time in iterations	1743	88
Mean recurrence time in seconds	10.39	2.14
Mean recurrence time in posterior density evaluations	15,862	10,315

Table 2: General diagnostics for the SP500 dataset; The acceptance rates for the first untempered chain are reported

		Posterior odds standard errors	
Model	True odds	Population reversible jump	Population SCC
2	0.8396	0.2175	0.0497
3	0.4953	0.0234	0.0818
4	0.2425	0.0100	0.0135
5	0.2044	0.0209	0.0170
6	0.1977	0.0175	0.0174
7	0.1680	0.0289	0.0250
8	0.1417	0.0059	0.0039
9	0.1044	0.0170	0.0045
10	0.1017	0.0053	0.0071
MSE		0.0674	0.0309

Table 3: Posterior odds standard errors for the SP500 dataset

probability model includes Applied Materials Inc., Computer Sciences Corp., Chevron Corp., Danaher Corp., IAC/Interactive Corp., New York Times Co. and Schwab (Charles) Corp. It is interesting to note that these stocks have neither the most shares outstanding (shares that have been authorized, issued, and purchased by investors) nor the highest capitalisation (share price times the shares outstanding) and they belong to different industry groups. Thus the model selection exercise identifies a collection of stocks with rather differentiated characteristics. In Table 4 we present the list of the most important stocks according to their marginal probability of inclusion, ordered according to their marginal probability in population SCC Algorithm.

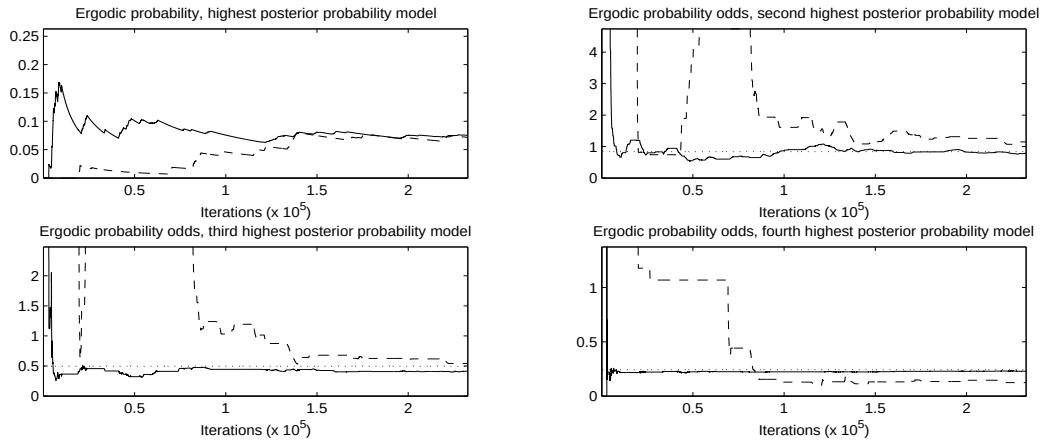


Figure 1: Ergodic probability and probability odds for the four highest probability models; SP500 dataset

5 Gene selection

We consider the problem of gene selection based on microarray expression data. The goal is to identify the responsible genes associated with the presence of hereditary breast cancer. The problem lies on the availability of a small sample size with a large number of predictors (genes). This causes extreme conditions of uncertainty which results to a challenging model determination problem. We use logistic regression and adopt the population reversible jump and the population SCC employed in the previous Section to implement the Bayesian analysis.

Our dataset has been used in Hedenfalk et al. (2001) and consists of 22 observations and 3226 different genes. The 22 breast tumor samples were taken from 21 patients. From these samples, 7, 8 and 7 had BRCA1-related, BRCA2-related and sporadic cancer respectively. Interest lies in identifying the genes associated with the presence of BRCA1-related cancer, versus the genes associated with BRCA2-related or sporadic cancer. BRCA1 and BRCA2 are proteins that participate in DNA repair and homologous recombination. A cell that lacks functional BRCA1 or BRCA2 protein, has decreased ability to repair damaged DNA. However certain pathological features can help to distinguish tumors with BRCA1 mutations from those with BRCA2 and sporadic mutations. For BRCA1 there is a higher meiotic count, pushing tumor margins and a lumphocytic infiltrate, while these mutations are generally negative for both estrogen and progesterone receptors. On the other hand BRCA2 mutations

Stock name	Marginal probability	
	Population reversible jump	Population SCC
Computer Sciences Corp.	0.9611	0.9562
New York Times Co.	0.7177	0.6918
Embarq Corp.	0.5126	0.5044
Danaher Corp.	0.3621	0.3658
Bristol-Myers Squibb Co.	0.4026	0.3505
Sunoco Inc.	0.3032	0.3085
Federal Investors Inc.	0.3144	0.2760
Chevron Corp.	0.2664	0.2680
Applied Materials Inc.	0.2338	0.2446
IAC/Interactive Corp.	0.1893	0.2010

Table 4: Marginal probability of inclusion; SP500 dataset

are heterogeneous and display substantially less tubule formation, while most of them are positive for hormone receptors.

The observations in this dataset refer to gene expression ratios, derived from the fluorescent intensity (proportional to the gene expression level) from a tumor sample (BRCA1, BRCA2, or sporadic), divided by the fluorescent intensity from a common reference sample (MCF-10A cell line). The common reference sample is used for all 21 microarray experiments. Therefore, the ratio may take values in $(0, \infty)$. In most studies analyzing gene expression ratios, a logarithm-transform is performed to convert the ratio data in order to achieve the symmetric property from over-expression to under-expression range. The same approach is followed in this application. Furthermore the original data have been truncated from below at 0.1 and above at 20 as in Zhou et al. (2004a).

Three studies that analyse this dataset in the context of Bayesian variable selection, are those of Lee et al. (2003); Zhou et al. (2004a) and Zhou et al. (2004b). In Lee et al. (2003) and Zhou et al. (2004b) a probit regression model is adopted to find the most important genes, while in Zhou et al. (2004a) a logistic regression model which is approximated by a normal linear model is used. In these studies the parameters of the linear predictor are integrated out and model selection is performed in model space. In our analysis we run our challenging model determination MCMC algorithm without performing a parameter integration, so samples from both parameters and models are obtained.

The prior distribution for the parameters of the logistic regression are taken to be multivariate normal following the construction of Ntzoufras et al. (2003). This prior specification corresponds to the unit information prior which results to a prior equivalent to that of a 'single prior observation'. For comparison purposes we follow Zhou et al. (2004b) and adopt their prior model probability for model size which is binomial truncated at 20 with parameter 5/3226. This encourages a priori parsimonious model specifications.

We used 5 chains in our population based algorithms both because the untempered chain

turned out to be quite mobile and because when we tried to include more than 5 chains we observed that the highly tempered chains visited mainly models with $19 - 20$ predictors. The temperature ladder is specified as in Section 3, with $z = 0.8852$.

The pseudoprior densities for both algorithms are taken to be normal distributions with mean and variance obtained as follows. We ran pilot MCMC runs of logistic regressions with only one variable using random walk Metropolis sampling. By running this algorithm 3226 times (for each gene), for $10,000$ iterations after a burn-in of 1000 iterations, we obtained ergodic sample means and variances that we used to construct our pseudoprior densities. In fact, we doubled the ergodic sample standard deviation since uncertainty was expected to increase in models with up to 20 variables. The update of the parameters within each model is performed via a multivariate random walk Metropolis with covariance matrix taken as the maximum likelihood estimate appropriately scaled.

We performed a similar study as in Section 4 which is governed by our desire to compare the two algorithms when there is equal number of posterior densities evaluations. The general diagnostics for the algorithms are presented in Table 5. The mean recurrence time of population reversible jump turns to be smaller than that of the population SCC in terms of CPU-time and posterior density evaluations.

	Population reversible jump	Population SCC
Sample taken	1,000,000	377,352
Overall time (seconds)	249,369	166,508
Overall evaluations	1,257,612,852	1,256,156,871
Average acceptance rate of addition moves	0.3096	0.5997
Average acceptance rate of deletion moves	0.3111	0.6037
Average acceptance rate of replacement moves	0.1860	0.1891
Average acceptance rate of update parameters moves	0.2159	0.2189
Average acceptance rate of exchange moves	0.4912	0.4829
Average acceptance rate of crossover moves	0.0277	0.0275
Mean recurrence time in iterations	2615.99	1681.03
Mean recurrence time in seconds	608.18	691.57
Mean recurrence time in posterior density evaluations	3,067,169	5,217,274

Table 5: General diagnostics for the genes dataset; The acceptance rates for the first untempered chain are reported

In Table 6 the posterior mean standard errors of the estimated probability odds with respect to the approximate odds, calculated with a Laplace approximation (see DiCiccio et al., 1997), for models with highest posterior probability are presented. The population SCC algorithm has significantly lower mean standard error than the population reversible jump. In Figure 2 the ergodic probability for the highest posterior probability model and the ergodic probability odds for the next three models against the first are displayed. In general the population SCC seems to approach closer and faster the Laplace-based approximate posterior odds.

Model	Approximate odds	Posterior odds standard errors	
		Population reversible jump	Population SCC
2	0.7110	0.4116	0.1463
3	0.2355	0.0238	0.0116
4	0.2144	0.0624	0.0821
5	0.1257	0.0950	0.0743
6	0.1215	0.0684	0.0936
7	0.1192	0.1247	0.0060
8	0.1086	0.0231	0.0149
9	0.1025	0.0530	0.0857
10	0.0904	0.0413	0.0743
MSE		0.1960	0.0891

Table 6: Posterior odds standard errors for the genes dataset; Approximate posterior odds are calculated with Laplace approximation

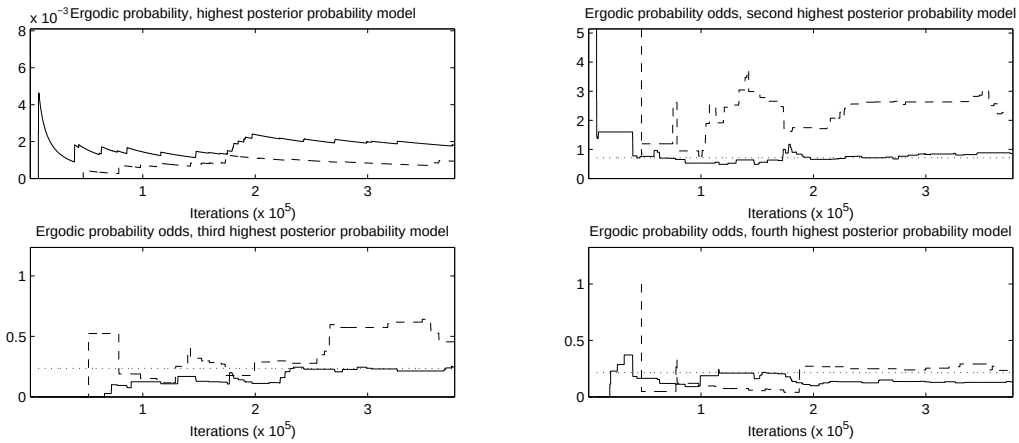
The list of the 10 models with the highest posterior model probability, along with their estimated model probability and parameters are presented in Table 7. All these models include only a single variable, although both algorithms visit frequently models with more variables as displayed in Figure 3. It is encouraging that the population SCC sorted in the right way the three models with the highest posterior model probability, while 8 out of these 10 models turned out to belong in the list of 10 top models with highest estimated posterior model probability.

Model	Genes	Population reversible jump			Population SCC		
		Est. Prob. (×100)	Coef.	Coef. std.	Est. Prob. (×100)	Coef.	Coef. std.
1	336	0.143	-5.0723	1.6740	0.180	-5.1363	1.7540
2	2300	0.160	9.6155	3.2959	0.154	8.9983	2.8920
3	118	0.037	7.9263	2.4422	0.045	7.4664	2.3476
4	10	0.040	3.5103	1.1166	0.024	3.1995	1.3136
5	955	0.032	2.4984	0.9203	0.036	2.5701	0.9554
6	1120	0.027	-7.5645	2.6433	0.005	-7.9902	3.192
7	560	0.035	5.1638	1.7448	0.020	4.6152	1.439
8	3219	0.019	-3.4173	1.2303	0.022	-3.1893	1.0425
9	1503	0.022	-3.4696	1.3090	0.034	-3.7592	1.1562
10	2321	0.007	10.7808	4.0112	0.030	11.3673	3.8905

Table 7: List of the 10 best models, estimated posterior probability and parameters estimates; Reported is the index of each gene in the order included in the dataset

When predictions are made based on the results in Table 7 under a model averaging basis, population SCC has no misclassifications whereas population reversible jump has one misclassification. These results are similar to other studies analysing this dataset.

In Table 8 the genes are classified according to their posterior marginal probability of inclusion and for comparison purposes we report significant genes found by relevant studies analysing the same dataset. It is interesting to note that only genes 336, 10, 118 and 1859



Dashed: Population reversible jump, Solid: Population SCC, Dotted: True odds

Figure 2: Ergodic probability and probability odds for the four highest probability models; Genes dataset

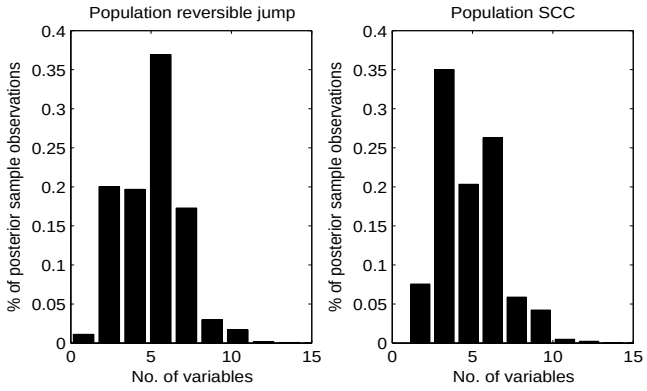


Figure 3: Number of variables included in the models present in the posterior sample; Genes dataset

are found to be significant in all three studies. Out of the top twenty genes found by the population SCC algorithm, 11 of them have been identified in Lee et al. (2003), 9 in Zhou et al. (2004a) and 11 in Zhou et al. (2004b).

	Population reversible jump		Population SCC		Lee et al. (2003)	Zhou et al. (2004a)	Zhou et al. (2004b)
No.	Gene	Prob./Prior	Gene	Prob./Prior	Gene	Gene	Gene
1	1999	26.89	1999	27.17	1008	10	10
2	2300	20.69	2423	23.57	336	118	336
3	10	19.18	1443	21.64	10	336	858
4	1008	16.51	336	19.21	1068	2699	733
5	336	15.05	2300	18.06	496	2761	2423
6	2222	14.13	2734	15.98	118	742	955
7	955	13.70	2222	15.47	3009	2382	1443
8	1443	13.50	955	13.57	585	2018	1417
9	2734	12.96	1859	13.09	523	157	2428
10	2423	12.18	10	12.95	556	739	118
11	858	11.22	1277	10.39	1999	1120	496
12	408	10.47	1008	10.34	2423	2272	1120
13	2699	9.99	560	10.24	498	1620	2018
14	1859	9.56	408	9.89	140	1999	523
15	118	8.56	118	9.59	1277	1859	1766
16	560	8.13	2428	9.34	955	439	2699
17	2259	8.09	496	9.14	272	2734	1859
18	585	7.80	2699	8.24	2734	247	1008
19	3199	7.65	2226	8.15	1859	3009	1179
20	2226	7.16	742	7.85	555	2423	555

Table 8: List of the most significant genes according to their marginal probability of inclusion as found by Population reversible jump, Population SCC and other studies; Reported is the index of each gene in the order included in the dataset; Prob./Prior is the ratio of the marginal probability of inclusion of each gene over the prior

6 Concluding remarks

The SCC algorithm is a simple modification of existing algorithms that has performed rather well in a series of simulated and real data examples. We feel that its simplicity and its efficiency offer a valuable tool for Bayesian model determination regression problems with large p.

References

Bottolo, L. and Richardson, S. (2010), "Evolutionary Stochastic Search for Bayesian Model Exploration," *Bayesian Analysis*, 5, 583–618.

Carlin, B. P. and Chib, S. (1995), "Bayesian model choice via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society, Series B*, 57, 473–484.

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

Dellaportas, P. and Roberts, G. O. (2003), *An introduction to MCMC*, New York: Springer-Verlag, Spatial Statistics and Computational methods, pp. 1–42.

DiCiccio, T., Kass, R., and Wasserman, L. (1997), "Computing Bayes factors by combining simulation and asymptotic approximations," *Journal of the American Statistical Association*, 92, 903–915.

Fernandez, C., Ley, E., and Steel, M. F. J. (2001), "Benchmark priors for Bayesian model averaging," *Journal of Econometrics*, 75, 317–343.

George, E. I. and McCulloch, R. E. (1997), "Approaches for Bayesian Variable selection," *Statistica Sinica*, 7, 339–373.

Godsill, S. J. (2001), "On the Relationship between Markov Chain Monte Carlo Methods for Model Uncertainty," *Journal of Computational and Graphical Statistics*, 10, 230–248.

Green, P. J. (1995), "Reversible jump MCMC computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Grimmett, G. and Stirzaker, D. (2004), *Probability and Random Processes*, Oxford University Press, 3rd ed.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for 'Large p' Regression," *Journal of the American Statistical Association*, 478, 507–516.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallionemi, O., Wilfond, B., Borg, A., and Trent, J. (2001), "Gene-expression profiles in hereditary breast cancer," *The New England Journal of Medicine*, 344, 539–548.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007b), "Population-based reversible jump Markov chain Monte Carlo," *Biometrika*, 94, 787–807.

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric regression using linear combinations of basis functions," *Statistics and Computing*, 11, 313–322.

Lee, K. E., Sha, N., Dougherty, E. R., Vanucci, M., and Mallick, B. K. (2003), "Gene selection: a Bayesian variable selection approach," *Bioinformatics*, 19, 90–97.

Ley, E. and Steel, M. F. J. (2009), "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression," *Journal of Applied Econometrics*, 24, 651–674.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer.

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

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003), "Bayesian variable and link determination for generalised linear models," *Journal of Statistical Planning and Inference*, 111, 165–180.

O'Hagan, A. and Forster, J. (2004), *Kendall's Advanced Theory of Statistics. Vol. 2B: Bayesian Inference*, New York: Oxford University Press Inc., 2nd ed.

Petralias, A. (2010), "Bayesian Model Determination and Nonlinear Threshold Volatility Models," Ph.D. thesis, Department of Statistics, Athens University of Economics and Business, Greece.

Zhou, X., Liu, K., and Wong, S. T. C. (2004a), "Cancer classification and prediction using logistic regression with Bayesian gene selection," *Journal of Biomedical Informatics*, 37, 249–259.

Zhou, X., Wang, X., and Dougherty, E. R. (2004b), "A Bayesian approach to nonlinear probit gene selection and classification," *Journal of The Franklin Institute*, 341, 137–156.