

# Evolutionary Stochastic Search

Leonardo Bottolo

Institute for Mathematical Sciences, Imperial College London, UK

[l.bottolo@imperial.ac.uk](mailto:l.bottolo@imperial.ac.uk)

Sylvia Richardson\*

Centre for Biostatistics, Imperial College, London, UK

[sylvia.richardson@imperial.ac.uk](mailto:sylvia.richardson@imperial.ac.uk)

## Abstract

Implementing Bayesian variable selection for linear Gaussian regression models for analysing high dimensional data sets is of current interest in many fields. In order to make such analysis operational, we propose a new search algorithm based upon Evolutionary Monte Carlo and designed to work equally well when  $n > p$  or under the “large  $p$ , small  $n$ ” paradigm, thus making multivariate analysis feasible, for example, in genetics/genomics experiments. Two real data examples in genomics are presented, demonstrating the performance of the algorithm in a space of up to 10,000 covariates. Finally the methodology is compared with a recently proposed search algorithms in an extensive simulation study.

*Keywords:* Evolutionary Monte Carlo; Fast Scan Metropolis-Hastings schemes; Linear Gaussian regression models; Variable selection.

---

\*Address for correspondence: Sylvia Richardson, Department of Epidemiology and Public Health, Imperial College, 1 Norfolk Place, London, W2 1PG, UK.

# 1 Introduction

This paper is a contribution to the methodology of Bayesian variable selection for linear Gaussian regression models, an important problem which has been much discussed both from a theoretical and a practical perspective (see Chipman *et al.*, 2001 and Clyde and George, 2004 for literature reviews). Recent advances have been made in two directions, unravelling the theoretical properties of different choices of prior structure for the regression coefficients (Fernández *et al.*, 2001; Liang *et al.*, 2008) and proposing algorithms that can explore the huge model space consisting of all the possible subsets when there are a large number of covariates, using either MCMC or other search algorithms (Kohn *et al.*, 2001; Dellaportas *et al.*, 2002; Hans *et al.*, 2007).

In this paper, we propose a new sampling algorithm for implementing the variable selection model, based on tailoring ideas from Evolutionary Monte Carlo (Liang and Wong, 2000; Jasra *et al.*, 2007) in order to overcome the known difficulties that MCMC samplers face in a high dimension multimodal model space: enumerating the model space becomes rapidly unfeasible even for a moderate number of covariates. For a Bayesian approach to be operational, it needs to be accompanied by an algorithm that samples the indicators of the selected subsets of covariates, together with any other parameters that have not been integrated out. Our new algorithm for searching through the model space has many generic features that are of interest *per se* and can be easily coupled with any prior formulation for the variance-covariance of the regression coefficients. We illustrate this by implementing  $g$ -priors for the regression coefficients as well as independent priors: in both cases the formulation we adopt is general and allows the specification of a further level of hierarchy on the priors for the regression coefficients, if so desired.

The paper is structured as follows. In Section 2, we present the background of Bayesian variable selection, reviewing briefly alternative prior specifications for the regression coefficients, namely  $g$ -priors and independent priors. Section 3 is devoted to the description of our MCMC sampler which uses a wide portfolio of moves. Section 4 demonstrates the good performance of our new MCMC algorithm in a variety of real and simulated examples with different structures on the predictors. In Section 4.2 we complement the results of the simulation study by comparing our algorithm with the recent Shotgun

Stochastic Search algorithm of Hans *et al.* (2007). Finally Section 5 contains some concluding remarks.

## 2 Background

### 2.1 Variable selection

Let  $y = (y_1, \dots, y_n)^T$  be a sequence of  $n$  observed responses and  $x_i = (x_{i1}, \dots, x_{ip})^T$  a vector of predictors for  $y_i$ ,  $i = 1, \dots, n$ , of dimension  $p \times 1$ . Moreover let  $X$  be the  $n \times p$  design matrix with  $i$ th row  $x_i^T$ . A Gaussian linear model can be described by the equation

$$y = \alpha 1_n + X\beta + \varepsilon,$$

where  $\alpha$  is an unknown constant,  $1_n$  is a column vector of ones,  $\beta = (\beta_1, \dots, \beta_p)^T$  is a  $p \times 1$  vector of unknown parameters and  $\varepsilon \sim N(0, \sigma^2 I_n)$ .

Suppose one wants to model the relationship between  $y$  and a subset of  $x_1, \dots, x_p$ , but there is uncertainty about which subset to use. Following the usual convention of only considering models that have the intercept  $\alpha$ , this problem, known as variable selection or subset selection, is particularly interesting when  $p$  is large and parsimonious models containing only a few predictors are sought to gain interpretability. From a Bayesian perspective the problem is tackled by placing a constant prior density on  $\alpha$  and a prior on  $\beta$  which depends on a latent binary vector  $\gamma = (\gamma_1, \dots, \gamma_p)^T$ , where  $\gamma_j = 1$  if  $\beta_j \neq 0$  and  $\gamma_j = 0$  if  $\beta_j = 0$ ,  $j = 1, \dots, p$ . The overall number of possible models defined through  $\gamma$  grows exponentially with  $p$  and selecting the best model that predicts  $y$  is equivalent to find one over the  $2^p$  subsets that form the model space.

Given the latent variable  $\gamma$ , a Gaussian linear model can therefore be written as

$$y = \alpha 1_n + X_\gamma \beta_\gamma + \varepsilon, \tag{1}$$

where  $\beta_\gamma$  is the non-zero vector of coefficients extracted from  $\beta$ ,  $X_\gamma$  is the design matrix of dimension  $n \times p_\gamma$ ,  $p_\gamma \equiv \gamma^T 1_p$ , with columns corresponding to  $\gamma_j = 1$ . We will assume that, apart from the intercept  $\alpha$ ,  $x_1, \dots, x_p$  contains no variables that would be included in every possible model and that the columns of the design matrix have all been centred with mean 0.

It is recommended to treat the intercept separately and assign it a constant prior:  $p(\alpha) \propto 1$ , Fernández *et al.* (2001). When coupled with the latent variable  $\gamma$ , the conjugate prior structure of  $(\beta_\gamma, \sigma^2)$  follows a normal-inverse-gamma distribution

$$p(\beta_\gamma | \gamma, \sigma^2) = N(m_\gamma, \sigma^2 \Sigma_\gamma) \quad (2)$$

$$p(\sigma^2 | \gamma) = p(\sigma^2) = \text{InvGa}(a_\sigma, b_\sigma) \quad (3)$$

with  $a_\sigma, b_\sigma > 0$ . Some guidelines on how to fix the value of the hyperparameters  $a_\sigma$  and  $b_\sigma$  are provided in Kohn *et al.* (2001), while the case  $a_\sigma = b_\sigma = 0$  corresponds to the Jeffreys' prior for the error variance,  $p(\sigma^2) \propto \sigma^{-2}$ . Taking into account (1), (2), (3) and the prior specification for  $\alpha$ , the joint distribution of all the variables (based on further conditional independence conditions) can be written as

$$p(y, \gamma, \alpha, \beta_\gamma, \sigma^2) = p(y | \gamma, \alpha, \beta_\gamma, \sigma^2) p(\alpha) p(\beta_\gamma | \gamma, \sigma^2) p(\sigma^2) p(\gamma). \quad (4)$$

The main advantage of the conjugate structure (2) and (3) is the analytical tractability of the marginal likelihood whatever the specification of the prior covariance matrix  $\Sigma_\gamma$ :

$$\begin{aligned} & \int p(y | \gamma, \alpha, \beta_\gamma, \sigma^2) p(\alpha) p(\beta_\gamma | \gamma, \sigma^2) p(\sigma^2) d\alpha d\beta_\gamma d\sigma^2 \\ & \propto |X_\gamma^T X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (2b_\sigma + S(\gamma))^{-(2a_\sigma + n - 1)/2}, \end{aligned} \quad (5)$$

where  $S(\gamma) = C - M^T K_\gamma^{-1} M$ , with  $C = (y - \bar{y}_n)^T (y - \bar{y}_n) + m_\gamma^T \Sigma_\gamma^{-1} m_\gamma$ ,  $M = X_\gamma^T (y - \bar{y}_n) + \Sigma_\gamma^{-1} m_\gamma$  and  $K_\gamma = X_\gamma^T X_\gamma + \Sigma_\gamma^{-1}$  (Brown *et al.*, 1998).

While the mean of the prior (2) is usually set equal to zero,  $m_\gamma = 0$ , a neutral choice (Chipman *et al.*, 2001; Clyde and George, 2004), the specification of the prior covariance  $\Sigma_\gamma$  matrix leads to at least two different classes of priors:

- When  $\Sigma_\gamma = gV_\gamma$ , where  $g$  is a scalar and  $V_\gamma = (X_\gamma^T X_\gamma)^{-1}$ , it replicates the covariance structure of the likelihood giving rise to so called  $g$ -priors first proposed by Zellner (1986).
- When  $\Sigma_\gamma = cV_\gamma$ , but  $V_\gamma = I_{p_\gamma}$  the components of  $\beta_\gamma$  are conditionally independent and the posterior covariance matrix is driven towards the independence case.

We will adopt the notation  $\Sigma_\gamma = \tau V_\gamma$  as we want to cover both prior specification in a unified manner. Thus in the  $g$ -prior case,  $\Sigma_\gamma = \tau (X_\gamma^T X_\gamma)^{-1}$  while in the independent case,  $\Sigma_\gamma = \tau I_{p_\gamma}$ . We will refer to  $\tau$  as the *variable selection coefficient* for reasons that will become clear in the next Section.

To complete the prior specification in (4),  $p(\gamma)$  must be defined. A complete discussion about alternative priors on the model space can be found in Chipman (1996) and Chipman *et al.* (2001). Here we adopt the beta-binomial prior illustrated in Kohn *et al.* (2001)

$$p(\gamma) = \int p(\gamma|\omega) p(\omega) d\omega = \frac{B(p_\gamma + a_\omega, p - p_\gamma + b_\omega)}{B(a_\omega, b_\omega)} \quad (6)$$

with  $p_\gamma \equiv \gamma^T 1_p$ , where the choice  $p(\gamma|\omega) = \omega^{p_\gamma} (1 - \omega)^{p - p_\gamma}$  implicitly induces a binomial prior distribution over the model size and  $p(\omega) = \omega^{a_\omega - 1} (1 - \omega)^{b_\omega - 1} / B(a_\omega, b_\omega)$ . The hypercoefficients  $a_\omega$  and  $b_\omega$  can be chosen once  $E(p_\gamma)$  and  $V(p_\gamma)$  have been elicited. In the “large  $p$ , small  $n$ ” framework, to ensure sparse regression models where  $p_\gamma \ll p$ , it is recommended to centre the prior for the model size away from the number of observations.

## 2.2 Priors for the variable selection coefficient $\tau$

### 2.2.1 $g$ -priors

It is a known fact that  $g$ -priors have two attractive properties. Firstly they possess an automatic scaling feature (Chipman *et al.*, 2001; Kohn *et al.*, 2001). In contrast, for independent priors, the effect of  $V_\gamma = I_{p_\gamma}$  on the posterior distribution depends on the relative scale of  $X$  and standardisation of the design matrix to units of standard deviation is recommended. However, this is not always the best procedure when  $X$  is possibly skewed, or when the columns of  $X$  are not defined on a common scale of measurement. The second feature that makes  $g$ -priors particularly appealing is the rather simple structure of the marginal likelihood (5) with respect to the constant  $\tau$  which becomes

$$\propto (1 + \tau)^{-p_\gamma/2} (2b_\sigma + S(\gamma))^{-(2a_\sigma + n - 1)/2}, \quad (7)$$

where, if  $m_\gamma = 0$ ,  $S(\gamma) = (y - \bar{y}_n)^T (y - \bar{y}_n) - \frac{\tau}{1 + \tau} (y - \bar{y}_n)^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T (y - \bar{y}_n)$ . For computational reasons explained in the next Section, we assume that (7) is always defined: since we calculate  $S(\gamma)$  using the QR-decomposition of the regression  $(X_\gamma, y - \bar{y}_n)$  (Brown *et al.*, 1998), when

$n \leq p_\gamma$ ,  $S(\gamma) = (y - \bar{y}_n)^T (y - \bar{y}_n) / (1 + \tau)$ . Despite the simplicity of (7), the choice of the constant  $\tau$  for  $g$ -priors is complex, see Fernández *et al.* (2001), Cui and George (2008) and Liang *et al.* (2008).

Historically the first attempt to build a comprehensive Bayesian analysis placing a prior distribution on  $\tau$  dates back to Zellner and Siow (1980), where the data adaptivity of the degree of shrinkage adapts to different scenarios better than assuming standard fixed values. Zellner-Siow priors, Z-S hereafter, can be thought as a mixture of  $g$ -priors and an inverse-gamma prior on  $\tau$ ,  $\tau \sim InvGa(1/2, n/2)$ , leading to

$$p(\beta_\gamma | \gamma, \sigma^2) \propto \int N\left(0, \sigma^2 \tau (X_\gamma^T X_\gamma)^{-1}\right) p(\tau) d\tau. \quad (8)$$

Liang *et al.* (2008) analyse in details Z-S priors pointing out a variety of theoretical properties. From a computational point of view, with Z-S priors, the marginal likelihood  $p(y | \gamma) = \int p(y | \gamma, \tau) p(\tau) d\tau$  is no more available in closed form, something which is advantageous in order to quickly perform a stochastic search (Chipman *et al.*, 2001). Even though Z-S priors need no calibration and the Laplace approximation can be derived (Tierney and Kadane, 1986), see Supplementary material (abbreviated as Suppl. mat.), Section A.2, they never became as popular as  $g$ -priors with a suitable constant value for  $\tau$ . For alternative priors, see also Cui and George (2008) and Liang *et al.* (2008).

## 2.2.2 Independent priors

When all the variables are defined on the same scale, independent priors represent an attractive alternative to  $g$ -priors. The likelihood marginalised over  $\alpha$ ,  $\beta_\gamma$  and  $\sigma^2$  becomes

$$p(y | \gamma) \propto \tau^{-p_\gamma/2} |X_\gamma^T X_\gamma + \tau I_{p_\gamma}|^{-1/2} (2b_\sigma + S(\gamma))^{-(2a_\sigma + n - 1)/2}, \quad (9)$$

where, if  $m_\gamma = 0$ ,  $S(\gamma) = (y - \bar{y}_n)^T (y - \bar{y}_n) - (y - \bar{y}_n)^T X_\gamma (X_\gamma^T X_\gamma + \tau I_{p_\gamma})^{-1} X_\gamma^T (y - \bar{y}_n)$ . Note that (9) is computationally more demanding than (7) due to the extra determinant operator.

Geweke (1996) suggests to fix a different value of  $\tau_j$ ,  $j = 1, \dots, p$ , based on the idea of “substantially significant determinant” of  $\Delta X_j$  with respect to  $\Delta y$ . However it is common practice to standardise the predictor variables, taking  $\tau = 1$  in order to place appropriate prior mass on reasonable values of the regression coefficients (Hans *et al.*, 2007). Another approach, illustrated in Bae and Mallick (2004), places a prior distribution on  $\tau_j$  without standardising the predictors (or on  $\tau$  after the standardisation).

Regardless of the prior specification for  $\tau$ , using the QR-decomposition on a suitable transformation of  $X_\gamma$  and  $y - \bar{y}_n$ , the marginal likelihood (9) is always defined.

### 3 MCMC sampler

In this Section we propose a new sampling algorithm that overcomes the known difficulties faced by MCMC schemes when attempting to sample a high dimension multimodal space. We discuss in a unified manner the general case where a hyperprior on the variable selection coefficient  $\tau$  is specified. This encompasses the  $g$ -prior and independent prior structure as well as the case of fixed  $\tau$  if a point mass prior is used.

The multimodality of the model space is a known issue in variable selection and several ways to tackle this problem have been proposed in the past few years. Liang and Wong (2000) suggest an extension of parallel tempering called Evolutionary Monte Carlo, EMC hereafter, Nott and Green, N&G hereafter, (2004) introduce a sampling scheme inspired by the Swendsen-Wang algorithm while Jasra *et al.* (2007) extend EMC methods to varying dimension algorithms. Finally Hans *et al.* (2007) propose when  $p > n$  a new stochastic search algorithm, SSS, to explore models that are in the same neighbourhood in order to quickly find the best combination of predictors.

We propose to solve the issue related to the multimodality of model space (and the dependence between  $\gamma$  and  $\tau$ ) along the lines of EMC, applying some suitable parallel tempering strategies directly on  $p(y|\gamma, \tau)$ . The basic idea of parallel tempering, PT hereafter, is to weaken the dependence of a function from its parameters by adding an extra one called “temperature”. Multiple Markov chains, called “population” of chains, are run in parallel, where a different temperature is attached to each chain, their state is tentatively swap at every sweep by a probabilistic mechanism and the latent binary vector  $\gamma$  of the non-heated chain is recorded. The different temperatures have the effect of flattening the likelihood. This ensures that the posterior distribution is not trapped in any local mode and that the algorithm mixes efficiently, since every chain constantly tries to transmit information about its state to the others. EMC extends this idea, encompassing the positive features of PT and genetic algorithms inside a MCMC scheme.

Since  $\beta$  and  $\sigma^2$  are integrated out, only two parameters need to be sampled, namely the latent binary vector and the variable selection coefficient. In this set-up the full conditionals to be considered are

$$[p(\gamma_l | \dots)]^{1/t_l} \propto [p(y | \gamma_l, \tau)]^{1/t_l} [p(\gamma_l)]^{1/t_l} \quad (10)$$

$$p(\tau | \dots) \propto \prod_{l=1}^L [p(y | \gamma_l, \tau)]^{1/t_l} p(\tau), \quad (11)$$

where  $L$  is the number of chains in the population and  $t_l$ ,  $1 = t_1 < t_2 < \dots < t_L$ , is the temperature attached to the  $l$ th chain while the population  $\gamma$  corresponds to a set of chains that are retained simultaneously. Conditions for convergence of EMC algorithms are well understood and illustrated for instance in Jasra *et al.* (2007).

At each sweep of our algorithm, first the population  $\gamma$  in (10) is updated using a variety of moves inspired by genetic algorithms: “local moves”, the ordinary Metropolis-Hastings or Gibbs update on every chain; and “global moves” that include: i) selection of the chains to swap, based on some probabilistic measures of distance between them; ii) crossover operator, i.e. partial swap of the current state between different chains; iii) exchange operator, full state swap between chains. Both local and global moves are important although global moves are crucial because they allow the algorithm to jump from one local mode to another. At the end of the update of  $\gamma$ ,  $\tau$  is then sampled using (11).

The implementation of EMC that we propose in this paper includes several novel aspects: the use of a wide range of moves including two new ones, a local move, based on the Fast Scan Metropolis-Hastings sampler, particularly suitable when  $p$  is large and a bold global move that exploits the pattern of correlation of the predictors. Moreover, we developed an efficient scheme for tuning the temperature placement. Another new feature is to use a Metropolis-within-Gibbs with adaptive proposal for updating  $\tau$ , as the full conditional (11) is not available in closed form.

### 3.1 EMC sampler for $\gamma$

In what follows, we will only sketch the rationale behind all the moves that we found useful to implement and discuss further the benefits of the new specific moves in Section 4.1. For the “large  $p$ , small  $n$ ” paradigm and complex predictor spaces, we believe that using a wide portfolio of moves is needed and

offers better guarantee of mixing.

From a notational point of view, we will use the double indexing  $\gamma_{l,j}$ ,  $l = 1, \dots, L$  and  $j = 1, \dots, p$  to denote the  $j$ th latent binary indicator in the  $l$ th chain. Moreover we indicate by  $\gamma_l = (\gamma_{l,1}, \dots, \gamma_{l,p})^T$  the vector of binary indicators that characterise the state of the  $l$ th chain of the population  $\gamma = (\gamma_1, \dots, \gamma_L)$ .

### Local moves and Fast Scan Metropolis Hastings sampler

Given  $\tau$ , we first implemented the simple MC<sup>3</sup> idea of Madigan and York (1995), also used by Brown *et al.* (1998) where add/delete and swap moves are used to update the latent binary vector  $\gamma_l$ . For an add/delete move, one of the  $p$  variables is selected at random and if the latent binary value is 0 the proposed new value is 1 or *vice versa*. However, when  $p \gg p_{\gamma_l}$ , where  $p_{\gamma_l}$  is the size of the current model for the  $l$ th chain, the number of sweeps required to select by chance a binary indicator with a value of 1 follows a geometric distribution with probability  $p_{\gamma_l}/p$  which is much smaller than  $1 - p_{\gamma_l}/p$  to select one with a value of 0. Hence, the algorithm spends most of the time trying to add rather than delete a variable. Note that this problem also affects RJ-type algorithms (Dellaportas *et al.*, 2002). On the other hand, Gibbs sampling (George and McCulloch, G&McC hereafter, 1993) is not affected by this issue since the state of the  $l$ th chain is updated by sampling from

$$[p(\gamma_{l,j} = 1 | y, \gamma_{l,j^-}, \tau)]^{1/t_l} \propto \exp \left\{ \left( \log p \left( y | \gamma_{l,j}^{(1)}, \tau \right) + \log p \left( \gamma_{l,j} = 1 | \gamma_{l,j^-} \right) \right) / t_l \right\}, \quad (12)$$

where  $\gamma_{l,j^-}$  indicates for the  $l$ th chain all the variables, but the  $j$ th,  $j = 1, \dots, p$  and

$\gamma_{l,j}^{(1)} = (\gamma_{l,1}, \dots, \gamma_{l,j-1}, \gamma_{l,j} = 1, \gamma_{l,j+1}, \dots, \gamma_{l,p})^T$ . The main problem related to Gibbs sampling is the large number of models it evaluates if a full Gibbs cycle or any permutation of the indices is implemented at each sweep. Each model requires the direct evaluation, or at least the update, of the time consuming quantity  $S(\gamma)$ , equation (7) or (9), making practically impossible to rely solely on the Gibbs sampler when  $p$  is very large. However, as sharply noticed by Kohn *et al.* (2001), it is wasteful to evaluate all the  $p$  updates in a cycle because if  $p_{\gamma_l}$  is much smaller than  $p$  and given  $\gamma_{l,j} = 0$ , it is likely that the sampled value of  $\gamma_{l,j}$  is again 0.

When  $p$  is large, we thus consider instead of the standard MC<sup>3</sup> add/delete, swap moves, two novel Fast Scan Metropolis-Hastings schemes, FSMH hereafter, specialised for EMC/PT. They are computationally less demanding than a full Gibbs sampling on all  $\gamma_{l,j}$  and do not suffer from the problem highlighted before for MC<sup>3</sup> and RJ-type algorithms when  $p \gg p_{\gamma_l}$ . The idea behind the FSMH move is to use an additional acceptance/rejection step (which is very fast to evaluate) to choose the number of indices where to perform the Gibbs-like step. One key point of our FSMH sampler is that the probability used in the acceptance/rejection step is flexible and based not only on the current chain model size  $p_{\gamma_l}$ , but also on the temperature  $t_l$  attached to the  $l$ th chain. Full details of the two FSMH schemes are given in the Appendix, while evaluation of them and comparison with MC<sup>3</sup> embedded in EMC are presented in Section 4.1.

### Global move: crossover operator

The first step of this move consists of selecting the pair of chains  $(l, r)$  to be operated on. We firstly compute a probability equal to the weight of the ‘‘Boltzmann probability’’,  $p_t(\gamma_l | \tau) = \exp\{f(\gamma_l | \tau)/t\}/F_t$ , where  $f(\gamma_l | \tau) = \log p(\gamma_l | y, \tau) + \log p(\gamma_l)$  is the log transformation of the full conditional (10) assuming  $t_l = 1 \forall l, l = 1, \dots, L$ , and  $F_t = \sum_{l=1}^L \exp\{f(\gamma_l | \tau)/t\}$  for some specific temperature  $t$ , and then rank all the chains according to this. We use normalised Boltzmann weights to increase the chance that the two selected chains will give rise, after the crossover, to a new configuration of the population with higher posterior probability. We refer to this first step as ‘‘selection operator’’.

Suppose that two new latent binary vectors are then generated from the selected chains according to some crossover operator described below. The new proposed population of chains

$\gamma' = (\gamma_1, \dots, \gamma'_l, \dots, \gamma'_r, \dots, \gamma_L)$  is accepted with probability

$$\alpha(\gamma \rightarrow \gamma') = \min \left\{ 1, \frac{\exp\{f(\gamma'_l | \tau)/t_l + f(\gamma'_r | \tau)/t_r\} Q_t(\gamma' \rightarrow \gamma | \tau)}{\exp\{f(\gamma_l | \tau)/t_l + f(\gamma_r | \tau)/t_r\} Q_t(\gamma \rightarrow \gamma' | \tau)} \right\}, \quad (13)$$

where  $Q_t(\gamma \rightarrow \gamma' | \tau)$  is the proposal probability, see Liang and Wong (2000).

In the following we will assume that four different crossover operators are selected at random at every EMC sweep: 1-point crossover, uniform crossover, adaptive crossover (Liang and Wong, 2000) and a

novel block crossover. Of these four moves, the uniform crossover which “shuffles” the binary indicators along all the selected chains is expected to have a low acceptance, but to be able to genuinely traverse regions of low posterior probability. The block crossover essentially tries to swap a group of variables that are highly correlated and can be seen as a multi-points crossover whose crossover points are not random but defined from the correlation structure of the covariates. In practice the block crossover is defined as follows: one variable is selected at random with probability  $1/p$ , then the pairwise correlation  $\rho(X_j, X_{j'})$  between the  $j$ th selected predictor and each of the remaining covariates,  $j' = 1, \dots, p$ ,  $j' \neq j$ , is calculated. We then retain for the block crossover all the covariates with positive (negative) pairwise correlation with  $X_j$  such that  $|\rho(X_j, X_{j'})| \geq \rho_0$ . The threshold  $\rho_0$  is chosen with consideration to the specific problem, but we fixed it at 0.25. Evaluation of block crossover and comparisons with other crossover operators are presented on a real data example in Section 4.1.

### **Global move: exchange operator**

The exchange operator can be seen as an extreme case of crossover operator, where the first proposed chain receives the whole second chain state  $\gamma'_l = \gamma_r$ , and *vice versa*. In order to achieve a good acceptance rate, the exchange operator is usually applied on adjacent chains in the temperature ladder, which limits its capacity for mixing. To obtain better mixing, we implemented two different approaches: the first one is based on Jasra *et al.* (2007) and the related idea of delayed rejection (Green and Mira, 2001); the second, a bolder “all-exchange” move, is based on a precalculation of all the  $L(L-1)/2$  exchange acceptance rates between all chains pairs (Calvo, 2005). Full relevant details are presented in Suppl. mat., Section A.1. Both of these bold moves perform well in the real data applications, see Section 4.1, and simulated examples, see Section 4.2, thus contributing to the efficiency of the algorithm.

### **Temperature placement**

As noted by Goswami and Liu (2007), the placement of the temperature ladder is the most important ingredient in population based MCMC methods. We propose a procedure for the temperature placement which has the advantage of simplicity while preserving good accuracy. First of all, we fix the size  $L$  of the population. In doing this, we are guided by several considerations: the complexity of the problem,

i.e.  $E(p_\gamma)$ , the size of the data and computational limits. We have experimented and we recommend to fix  $L \geq 3$ . Even though some of the simulated examples had  $p_\gamma \simeq 20$  (Section 4.2), we found that  $L = 5$  was sufficient to obtain good results. In our real data examples (Section 4.1), we used  $L = 4$  guided by some prior knowledge on  $E(p_\gamma)$ . Secondly, we fix at an initial stage, a temperature ladder according to a geometric scale such that  $t_{l+1}/t_l = b$ ,  $b > 1$ ,  $l = 1, \dots, L$  with  $b$  relatively large, for instance  $b = 4$ . To subsequently tune the temperature ladder, we then adopt a strategy based on monitoring only the acceptance rate of the delayed rejection exchange operator towards a target of 0.5. Details of the implementation are left to the Suppl. mat., Section A.1.

### 3.2 Adaptive Metropolis-within-Gibbs for $\tau$

Various strategies can be used to avoid having to sample from the posterior distribution of the variable selection coefficient  $\tau$ . The easiest way is to integrate it out through a Laplace approximation (Tierney and Kadane, 1986) or using a numerical integration such as quadrature on an infinite interval. We do not pursue these strategies and the reasons can be summarised as follows. Integrating out  $\tau$  in the population implicitly assumes that every chain has its own value of the variable selection coefficient  $\tau_l$  (and of the latent binary vector  $\gamma_l$ ). In this set-up, two unpleasant situations can arise: firstly, if a Laplace approximation is applied, *equilibrium* in the product space is difficult to reach because the posterior distribution of  $\gamma_l$  depends, through the marginal likelihood obtained using the Laplace approximation, on the *chain specific value* of the posterior mode for  $\tau_l$ ,  $\hat{\tau}_{\gamma_l}$ . Since the strength of  $X_{\gamma_l}$  to predict the response is weakened for chains attached to high temperatures, it turns out (details in Suppl. mat., Section A.2) that for these chains,  $\hat{\tau}_{\gamma_l}$  is likely to be close to zero. When the variable selection coefficient is very small, the marginal likelihood dependence on  $X_{\gamma_l}$  decreases even further, see for instance (7), and chains attached to high temperatures will experience a very unstable behaviour, making the convergence in the product space hard to reach. In addition, if an automatic tuning of temperature ladder is applied, chains will increasingly be placed at a closer distance in the temperature ladder to balance the low acceptance rate of the global moves, negating the purpose of EMC.

In this paper the convergence is reached instead in the product space  $\prod_{l=1}^L [p(\gamma_l | y, \tau)]^{1/t_l} p(\tau)$ , i.e.

the whole population is conditioned on a value of  $\tau$  *common to all chains*. This strategy will alleviate the problems highlighted before allowing for faster convergence and better mixing among the chains. The procedure just described comes with an extra cost, i.e. sampling the value of  $\tau$ . However, this step is inexpensive in relation to the cost required to sample  $\gamma_l$ ,  $l = 1, \dots, L$ . There are several strategies that can be used to sample  $\tau$  from (11). We found useful to apply the idea of adaptive Metropolis-within-Gibbs described in Roberts and Rosenthal (2008). Conditions for the asymptotic convergence and ergodicity are guaranteed as we enforce the *diminishing adaptive condition*, i.e. the transition kernel stabilises as the number of sweeps goes to infinity and the *bounded convergence condition*, i.e. the convergence time of the kernel is bounded in probability. In our set-up using an adaptive proposal to sample  $\tau$  has several benefits; amongst others it avoids the known problems faced by the Gibbs sampler when the prior is proper, but relatively flat (Natarajan and McCulloch, 1998) as can happen for Z-S priors when  $n$  is large or for the independent case considered by Bae and Mallick (2004). Moreover, given an upper limit on the number of sweeps, the adaptation guarantees a better exploration of the tails of  $p(\tau | y)$  than with a fixed proposal. For details of the implementation and discussion of conditions for convergence, see Suppl. mat., Section A.2.

### 3.3 ESS algorithm

In the following, we refer to our proposed algorithm, Evolutionary Stochastic Search as ESS. If  $g$ -priors are chosen the algorithm is denoted as ESS $g$ , while we use ESS $i$  if independent priors are selected (the same notation is used when  $\tau$  is fixed or given a prior distribution). Without loss of generality, we assume that the response vector and the design matrix have both been centred and, in the case of independent priors, that the design matrix is also rescaled. Based on the two full conditionals (10) and (11) and the local and global moves introduced earlier, our ESS algorithm can be summarised as follows.

- Given  $\tau$ , sample the population's states  $\gamma$  from the two steps:
  - (i) With probability 0.5 perform local move and with probability 0.5 apply at random one of the four crossover operators: 1-point, uniform, block and adaptive crossover. If local move is selected,

apply MC<sup>3</sup> if  $n > p$  and use FSMH sampling scheme 2 independently for each chain if  $p \geq n$  (see Appendix). Moreover when  $p \geq n$ , every 100 sweeps apply on the first chain a complete scan by a Gibbs sampler.

- (ii) Perform the delayed rejection exchange operator or the all-exchange operator with equal probability. During the burn-in, only select the delayed rejection exchange operator.
- When  $\tau$  is not fixed but has a prior  $p(\tau)$ , given the latent binary configuration  $\gamma = (\gamma_1, \dots, \gamma_L)$ , sample  $\tau$  from an adaptive Metropolis-within-Gibbs sampling (Section 3.2).

From a computational point of view, we used the same fast form for updating  $S(\gamma)$  as Brown *et al.* (1998), based on the QR-decomposition. Besides its numerical benefits, QR-decomposition can deal with the case  $p_\gamma \geq n$ . This avoids having to restrict the search to models with  $p_\gamma < n$ , and helps mixing during the burn-in phase.

## 4 Performance of ESS

### 4.1 Real data examples

The first real data example is an application of linear regression to investigate genetic regulation. To discover the genetic causes of variation in the expression (i.e. transcription) of genes, gene expression data are treated as a quantitative phenotype while genotype data (SNPs) are used as predictors, a type of analysis known as expression Quantitative Trait Loci (eQTL).

Here we focus on the ability of ESS to find a parsimonious set of predictors in an animal data set (Hubner *et al.*, 2005), where the number of observations,  $n = 29$ , is small with respect to the number of covariates  $p = 1,421$ . This situation, where  $n \ll p$ , is quite common in animal experiments since environmental sources of variation are controlled as well as the biological diversity of the sample. For illustration, we report the analysis of one gene expression response, where we apply ESS<sub>g</sub> with and without the hyperprior on  $\tau$ , see Table 1–eQTL. In the former case, thanks to the adaptive proposal, the Markov chain for  $\tau$  mixes very well reaching an overall acceptance rate which is close to the target value 0.44. Convergence issue is not a problem since the trace of the proposal’s standard deviation stabilises

quickly and well inside the bounded conditions, see Suppl. mat., Figure S.1.

In both cases a good mixing among the  $L = 4$  chains is obtained (Figure 1, top panels, ESSg with  $\tau = 29$ ). Although in the case depicted in Figure 1 with fixed  $\tau$ , the convergence is reached in the product space  $\prod_{l=1}^L [p(\gamma_l | y)]^{1/t_l}$ , by visual inspection we see that each chain *marginally* reaches its *equilibrium* with respect to the others; moreover, thanks to the automatic tuning of the temperature placement during the burn-in, the distributions of the chains log posterior probabilities overlap nicely, allowing effective exchange of information between the chains. Table 1–eQTL, confirms that the automatic temperature selection works well (with and without the hyperprior on  $\tau$ ) reaching an acceptance rate for the monitored exchange (delayed rejection) operator close to the selected target of 0.50. The all-exchange operator shows a higher acceptance rate, while, in contrast to Jasra *et al.* (2007), the overall crossover acceptance rate is reasonable high: in our experience the good performance of the crossover operator is both related to the selection operator (Section 3.1) and the new block crossover which shows an acceptance rate far higher than the others. Finally the computational time on the same desktop computer (see details in Suppl. mat., Section C.3) is rather similar with or without the hyperprior  $\tau$ , 28 and 30 minutes respectively for 25,000 sweeps with 5,000 as burn-in.

The main difference among the two implementations of ESSg is related to the posterior model size: when  $\tau$  is fixed at  $\tau = 29$  (Unit Information Prior, Fernández *et al.*, 2001), there is more uncertainty and support for larger models, see Figure 2 (a). In both cases we fix  $E(p_\gamma) = 4$  and  $V(p_\gamma) = 2$ , following prior biological knowledge on the genetic regulation. The posterior mean of the variable selection coefficient is a little smaller than the Unit Information Prior, with ESSg coupled with the Z-S prior favouring smaller models than when  $\tau$  is set equal to 29. The best model visited (and the corresponding  $R_\gamma^2 = 1 - S(\gamma)/y^T y$ ) is the same for both version of ESSg, while, when a hyperprior on  $\tau$  is implemented, the “stability index” which indicates how much the algorithm persists on the first chain top 1,000 (not unique) visited models ranked by the posterior probability (Suppl. mat., Section C.3), shows a higher stability, see Table 1–eQTL. In this case, having a data-driven level of shrinkage helps the search algorithm to better discriminate among competing models.

Our second example is related to the application of model (1) in another genomics example: 10,000 SNPs, selected genome-wide from a candidate gene study, are used to predict the variation of Mass Spectrography metabolomics data in a small human population, an example of a so-called mQTL experiment. A suitable dimension reduction of the data is performed to divide the spectra in regions or bins and  $\log_{10}$ -transformation is applied in order to normalise the signal.

We present the key findings related to a particular metabolite bin, but the same comments can be extended to the analysis of the whole data set, where we regressed every metabolites bin *versus* the genotype data ( $n = 50$  and  $p = 10,000$ ). In this very challenging case, we still found an efficient mixing of the chains (see Table 1–mQTL). Note that in this case the posterior mean of  $\tau$ , 63.577, is a little larger than the Unit Information Prior,  $\tau = n$ , although the influence of the hyperprior is less important than in the previous real data example, see Figure 2 (b). In both examples, the posterior model size favours clearly polygenic control with significant support for up to four genetic control points (Figure 2) highlighting the advantage of performing multivariate analysis in genomics rather than the traditional univariate analysis.

As expected in view of the very large number of predictors, in the mQTL example the computational time is quite large, around 5 hours for 20,000 sweeps after a burn-in of 5,000, but as shown in Table 1 by the “stability index” ( $\approx 0$ ), we believe that the number of iterations chosen exceeds what is required in order to visit faithfully the model space. For such large data analysis tasks, parallelisation of the code could provide big gains of computer time and would be ideally suited to our multiple chains approach.

[Table 1 about here – Figure 1 about here – Figure 2 about here]

Finally a referee asked us to show the superiority of our ESS algorithm, and in particular the FSMH schemes and the block crossover, with respect to more traditional EMC implementations illustrated for instance in Liang and Wong (2000). Albeit we believe that using a wide portfolio of different moves enables any searching algorithm to better explore complicated model spaces, we reanalysed the first real data example, eQTL analysis, comparing: (i) ESS $g$  with only FSMH $_2$  as local move *vs* ESS with only MC $^3$  as local move; (ii) ESS $g$  with only block crossover *vs* ESS $g$  with only 1-point, only uniform

and only adaptive crossover respectively. To avoid dependency of the results on the initialisation of the algorithm, we replicated the analysis 25 times. Moreover, to make the comparison fair, in experiment (i) we run the two versions of ESS $g$  for a different number of sweeps (25, 000 and 350, 000 with 5, 000 and 70, 000 as burn-in respectively), but matching the number of models evaluated at  $10^6$  for both. Results are presented in Suppl. mat., Table S.1. We report here the main findings:

- (i) over the 25 runs, ESS $g$  with FSMH $_2$  reaches the same top visited model 68% (17/25) while ESS $g$  with MC $^3$  the same top model only 28%, with a fixed  $\tau$ , and 88% and 40% respectively with Z-S prior. The great superiority when FSMH schemes are implemented can be explained by comparing subplot (a) and (c) in Figure 1: the exchange of information between chains for ESS $g$  with MC $^3$  as local move when  $p > n$  (and  $p \gg p_\gamma$ ) is rather poor, negating the purpose of EMC. ESS $g$  with MC $^3$  has more difficulties to reach convergence in the product space and, in contrast to ESS $g$  with FSMH $_2$ , the retained chain does not easily escape from local modes. This later point can be seen looking at Figure 1 (d) which magnifies the right hand tail of the kernel density of  $\log p(\gamma|y)$  for the recorded chain, pulling together the 25 runs: interestingly ESS $g$  with FSMH $_2$  is less “bumpy”, showing a better ability to escape from local modes and to explore more efficiently the right tail.
- (ii) Regarding the second comparison when  $\tau$  is fixed, ESS $g$  with only block crossover beats constantly the other crossover operators, with 80% vs about 60%, in terms of best model visited (Table S.1), has higher acceptance rate (Suppl. mat., Table S.2), showing also a great capacity to accumulate posterior mass as illustrated in Suppl. mat., Figure S.2. The specific benefit of the block crossover is less pronounced when a prior on  $\tau$  is specified, but we have already noticed that in this case having a hyperprior on  $\tau$  greatly improves the efficiency of the search.

## 4.2 Simulation study

We briefly report on a comprehensive study of the performance of ESS in a variety of simulated examples as well as a comparison with SSS. To make comparison with SSS fair, we use ESS $i$ , the version of our algorithm which assumes independent priors,  $\Sigma_\gamma = \tau I_{p_\gamma}$ , with  $\tau$  fixed at 1. Details of the simulated examples (6 set-ups) and how we conducted the simulation experiment (25 replication of each set-up) are

given in Suppl. mat., Section C. The rationale behind the construction of the examples was to benchmark our algorithm against both  $n > p$  and  $p > n$  cases, to use as building blocks intricate correlation structures that had been used in previous comparisons by G&McC (1993, 1997) and N&G (2004), as well as a realistic correlation structure derived from genetic data, and to include elements of model uncertainty in some of the examples by using a range of values of regression coefficients.

In our example we observe an effective exchange of information between the chains (reported in Suppl. mat., Table S.3) which shows good overall acceptance rates for the collection of moves that we have implemented. The dimension of the problem does not seem to affect the acceptance rates in Table S.3, remarkably since values of  $p$  range from 60 to 1,000 between the examples. We also studied specifically the performance of the global moves (Suppl. mat., Table S.4) to scrutinise our temperature tuning and confirmed the good performance of ESS*i* with good frequencies of swapping (not far from the case where adjacent chains are selected to swap at random with equal probability) and good measures of overlap between chains.

All the examples were run in parallel with ESS*i* and SSS 2.0 (Hans *et al.*, 2007) for the same number of sweeps (22,000) and matching hyperparameters on the model size. Comparison were made with respect to the marginal probability of inclusion as well as the ability to reach models with high posterior probability and to persist in this region. For a detailed discussion of all comparison, we refer the reader to Suppl. mat., Section C.3.

Overall the covariates with non-zero effects have high marginal posterior probability of inclusion for ESS*i* in all the examples (Suppl. mat., Figure S.4). There is good agreement between the two algorithms in general, with additional evidence on some examples (Figure S.4 (c) and (d)) that ESS*i* is able to explore more fully the model space and in particular to find small effects, leading to a posterior model size that is close to the true one. Measures of goodness of fit and stability, see Suppl. mat., Table S.5, are in good agreement between ESS*i* and SSS. The comparison highlight that a key feature of SSS, its ability to move quickly towards the right model and to persist on it, is accompanied by a drawback in having difficulty to explore far apart models with competing explanatory power, in contrast to ESS*i*

(contaminated example set-up). Altogether ESS $i$  shows a small improvement of  $R_{\gamma}^2$ , related to its ability to pick up some of the small effects that are missed by SSS. Finally ESS $i$  shows a remarkable superiority in terms of computational time, especially when the simulated (and estimated)  $p_{\gamma}$  is large.

## 5 Discussion

The key idea in constructing an effective MCMC sampler for  $\gamma$  and  $\tau$  is to add an extra parameter, the temperature, that weakens the likelihood contribution and enables escaping from local modes. Running parallel chains at different temperature is, on the other hand, expensive and the added computational cost has to be balanced against the gains arising from the various “exchanges” between the chains. This is why we focussed on developing a good strategy for selecting the pairs of chains, using both marginal and joint information between the chains, attempting bold and more conservative exchanges. Combining this with an automatic choice of the temperature ladder during burn-in is one of the key element of our ESS algorithm. Using PT in this way has the potential to be effective in a wide range of situations where the posterior space is multimodal.

To tackle the case where  $p$  is large with respect to  $p_{\gamma}$ , the second important element in our algorithm is the use of a Metropolised Gibbs sampling-like step performed on a subset of indices in the local updating of the latent binary vector, rather than an MC<sup>3</sup> or RJ-like updating move. The new Fast Scan Metropolis Hastings sampler that we propose to perform these local moves achieves an effective compromise between full Gibbs sampling that is not feasible at every sweep when  $p$  is large and vanilla add/delete moves. Detailed comparison of FSMH vs MC<sup>3</sup> scheme on a real data example shows the superiority of our new local move.

When a model with a prior on the variable selection coefficient  $\tau$  is preferred, the updating of  $\tau$  itself present no particular difficulties and is computationally inexpensive. Moreover, using an adaptive sampler makes the algorithm self contained without any time consuming tuning of the proposal variance. This latter strategy works perfectly well both in the  $g$ -prior and independent prior case as illustrated in Sections 4.1 and 4.2. Our current implementation does not make use of the output of the heated chains for posterior inference. Whether gains in variance reduction could be achieved in the spirit of Gramacy

*et al.* (2007) is an area for further exploration, which is beyond the scope of the present work.

Our approach has been applied so far to linear regression with univariate response  $y$ . An interesting generalisation is that of a multidimensional  $n \times q$  response  $Y$  and the identification of regressors that jointly predict the  $Y$  (Brown *et al.*, 1998). Much of our set-up and algorithm carries through without difficulties and we have already implemented our algorithm in this framework in a challenging case study in genomics with multidimensional outcomes.

## Acknowledgements

The authors are thankful to Norbert Hubner and Timothy Aitman for providing the data of the eQTL example, Gareth Roberts and Jeffrey Rosenthal for helpful discussions about adaptation and Michail Papatomas for his detailed comments. We are also grateful to the the editor, associate editor and three anonymous referees for valuable comments that greatly improved the presentation of the paper.

## Appendix

### A FSMH schemes

Let  $\gamma_{l,j}$ ,  $l = 1, \dots, L$  and  $j = 1, \dots, p$  to denote the  $j$ th latent binary indicator in the  $l$ th chain. As in Kohn *et al.* (2001), let  $\gamma_{l,j}^{(1)} = (\gamma_{l,1}, \dots, \gamma_{l,j-1}, \gamma_{l,j} = 1, \gamma_{l,j+1}, \dots, \gamma_{l,p})^T$  and

$\gamma_{l,j}^{(0)} = (\gamma_{l,1}, \dots, \gamma_{l,j-1}, \gamma_{l,j} = 0, \gamma_{l,j+1}, \dots, \gamma_{l,p})^T$ . Furthermore let  $L_{l,j}^{(1)} \propto p(y | \gamma_{l,j}^{(1)}, \tau)$  and  $L_{l,j}^{(0)} \propto p(y | \gamma_{l,j}^{(0)}, \tau)$  and finally  $\theta_{l,j}^{(1)} = p(\gamma_{l,j} = 1 | \gamma_{l,j^-})$  and  $\theta_{l,j}^{(0)} = 1 - \theta_{l,j}^{(1)}$ . From (6) it is easy to prove that

$$\theta_{l,j}^{(1)} = p(\gamma_{l,j} = 1 | \gamma_{l,j^-}) = \frac{p_{\gamma_l} + a_{\omega} - 1}{p + a_{\omega} + b_{\omega} - 1}, \quad (\text{A.1})$$

where  $p_{\gamma_l}$  is the current model size for the  $l$ th chain. Using the above equation, for  $\gamma_{l,j} = 1$  the normalised version of (12) can be written as

$$[p(\gamma_{l,j} = 1 | y, \gamma_{l,j^-}, \tau)]^{1/t_l} = \frac{\theta_{l,j}^{(1) 1/t_l} L_{l,j}^{(1) 1/t_l}}{S(1/t_l)}, \quad (\text{A.2})$$

where  $S(1/t_l) = \theta_{l,j}^{(1) 1/t_l} L_{l,j}^{(1) 1/t_l} + \theta_{l,j}^{(0) 1/t_l} L_{l,j}^{(0) 1/t_l}$  with  $[p(\gamma_{l,j} = 1 | y, \gamma_{l,j^-}, \tau)]^{1/t_l}$  defined similarly. Hence if  $\theta_{l,j}^{(1) 1/t_l}$  is very small, then  $[p(\gamma_{l,j} = 1 | y, \gamma_{l,j^-}, \tau)]^{1/t_l}$  is small as well. Therefore for the Gibbs sampler with a beta-binomial prior on the model space, the posterior probability of  $\gamma_{l,j} = 1$  depends crucially on  $\theta_{l,j}^{(1)}$ .

In the following we derive two Fast Scan Metropolis-Hastings schemes specialised for Evolutionary Monte Carlo or parallel tempering. We define  $Q(1 \rightarrow 0) = Q(\gamma_{l,j}^{(1)} \rightarrow \gamma_{l,j}^{(0)})$  as the proposal probability to go from 0 to 1 and  $Q(1 \rightarrow 0)$  the proposal probability to go from 1 to 0 for the  $j$ th variable and  $l$ th chain. Moreover using the notation introduced before, the Metropolis-within-Gibbs version of (12) to go from 0 to 1 in the EMC local move is

$$\alpha_l^{\text{MwG}}(0 \rightarrow 1) = \min \left\{ 1, \frac{\theta_{l,j}^{(1)1/t_l} L_{l,j}^{(1)1/t_l} Q(1 \rightarrow 0)}{\theta_{l,j}^{(0)1/t_l} L_{l,j}^{(0)1/t_l} Q(0 \rightarrow 1)} \right\} \quad (\text{A.3})$$

with a similar expression for  $\alpha_l^{\text{MwG}}(1 \rightarrow 0)$ . The proof of the Propositions are omitted since they are easy to check. We first introduce the following Proposition which is useful for the calculation of the acceptance probability in the FSMH schemes.

**Proposition 1** *The following three conditions are equivalent: a)  $L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} \geq 1$ ; b)  $L_{l,j}^{(1)1/t_l} / \tilde{S}(1/t_l) \geq 1$ ; c)  $L_{l,j}^{(0)1/t_l} / \tilde{S}(1/t_l) < 1$ , where  $\tilde{S}(1/t_l) = S(1/t_l) / \left( \theta_{l,j}^{(1)1/t_l} + \theta_{l,j}^{(0)1/t_l} \right)$  is the convex combination of the marginal likelihood  $L_{l,j}^{(1)1/t_l}$  and  $L_{l,j}^{(0)1/t_l}$  with weights  $\tilde{\theta}_{l,j}^{(1)}(1/t_l) = \theta_{l,j}^{(1)1/t_l} / \left( \theta_{l,j}^{(1)1/t_l} + \theta_{l,j}^{(0)1/t_l} \right)$  and  $\tilde{\theta}_{l,j}^{(0)}(1/t_l) = 1 - \tilde{\theta}_{l,j}^{(1)}(1/t_l)$ .*

**Proposition 2 (Sampling scheme 1)** *Let  $l = 1, \dots, L$ ,  $j = 1, \dots, p$  (or any permutation of them),*

$Q^{\text{FSMH}_1}(0 \rightarrow 1) = \tilde{\theta}_{l,j}^{(1)}(1/t_l) \min \left\{ 1, L_{l,j}^{(1)1/t_l} / \tilde{S}(1/t_l) \right\}$  and  
 $Q^{\text{FSMH}_1}(1 \rightarrow 0) = \tilde{\theta}_{l,j}^{(0)}(1/t_l) \min \left\{ 1, L_{l,j}^{(0)1/t_l} / \tilde{S}(1/t_l) \right\}$ . *Then the acceptance probabilities are*

$$\alpha_l^{\text{FSMH}_1}(0 \rightarrow 1) = \begin{cases} 1 & \text{if } L_{l,j}^{(1)1/t_l} / L_{l,j}^{(0)1/t_l} \geq 1 \\ \tilde{S}(1/t_l) / L_{l,j}^{(0)1/t_l} & \text{if } L_{l,j}^{(1)1/t_l} / L_{l,j}^{(0)1/t_l} < 1 \end{cases} \quad (\text{A.4})$$

$$\alpha_l^{\text{FSMH}_1}(1 \rightarrow 0) = \begin{cases} 1 & \text{if } L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} \geq 1 \\ \tilde{S}(1/t_l) / L_{l,j}^{(1)1/t_l} & \text{if } L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} < 1 \end{cases} \quad (\text{A.5})$$

The above sampling scheme is implemented as follows. For a given  $l$  and for  $j = 1, \dots, p$  (or any permutation of them) let  $u \sim U(0, 1)$ . Consider for simplicity  $0 \rightarrow 1$ . If  $u > \tilde{\theta}_{l,j}^{(1)}(1/t_l)$  then  $u > Q^{\text{FSMH}_1}(0 \rightarrow 1)$  and the move is rejected. If  $u \leq \tilde{\theta}_{l,j}^{(1)}(1/t_l)$  then  $Q^{\text{FSMH}_1}(0 \rightarrow 1)$  must be calculated

which is equivalent to evaluate (A.4). Sampling scheme 1 can be seen as a random scan Metropolis-within-Gibbs algorithm where the number of evaluations is linked to the prior/current model size and the temperature attached to the chain.

The second sampling scheme retains the idea of a two-step Metropolis-Hastings acceptance rate as in FSMH<sub>1</sub>. However it simplifies ever further the computation requirements using the normalised tempered version of (A.1) as a proposal.

**Proposition 3 (Sampling scheme 2)** Let  $l = 1, \dots, L$ ,  $j = 1, \dots, p$  (or any permutation of them),  $Q^{FSMH_2}(0 \rightarrow 1) = \tilde{\theta}_{l,j}^{(1)}(1/t_l)$  and  $Q^{FSMH_2}(1 \rightarrow 0) = \tilde{\theta}_{l,j}^{(0)}(1/t_l)$  with  $\tilde{\theta}_{l,j}^{(0)}(1/t_l) = 1 - \tilde{\theta}_{l,j}^{(1)}(1/t_l)$ .

The acceptance probabilities are

$$\alpha_l^{FSMH_2}(0 \rightarrow 1) = \begin{cases} 1 & \text{if } L_{l,j}^{(1)1/t_l} / L_{l,j}^{(0)1/t_l} \geq 1 \\ L_{l,j}^{(1)1/t_l} / L_{l,j}^{(0)1/t_l} & \text{if } L_{l,j}^{(1)1/t_l} / L_{l,j}^{(0)1/t_l} < 1 \end{cases} \quad (\text{A.6})$$

$$\alpha_l^{FSMH_2}(1 \rightarrow 0) = \begin{cases} 1 & \text{if } L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} \geq 1 \\ L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} & \text{if } L_{l,j}^{(0)1/t_l} / L_{l,j}^{(1)1/t_l} < 1 \end{cases} \quad (\text{A.7})$$

The above sampling scheme works as follows. Given the  $l$ th chain, if  $\gamma_{lj} = 0$  (and similarly for  $\gamma_{lj} = 1$ ), it proposes the new value from a Bernoulli distribution with probability  $\tilde{\theta}_{l,j}^{(1)}(1/t_l)$ : if the proposed value is different from the current one, it evaluates (A.6) otherwise it selects a new covariate.

Finally it can be proved that the Gibbs sampler is more efficient than the FSMH schemes, i.e. for a fixed number of iterations, Gibbs sampling MCMC standard error is lower than for FSMH samplers. However the Gibbs sampler is computationally more expensive so that, if  $p$  is very large, as described in Kohn *et al.* (2001), FSMH schemes become more efficient per floating point operation.

## References

- Bae, N. and Mallick, B.K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423-3430.
- Brown, P.J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc. B*, **60**, 627-641.

- Calvo, F. (2005) All-exchange parallel tempering. *J. Chem. Phys.*, **123**, 1-7.
- Chipman, H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.*, **24**, 17-36.
- Chipman, H., George, E.I. and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection (with discussion). In *Model Selection* (P. Lahiri, ed), 66-134. IMS: Beachwood, OH.
- Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Sci.*, **19**, 81-94.
- Cui, W. and George, E.I. (2008). Empirical Bayes vs fully Bayes variable selection. *J. Stat. Plan. Inf.*, **138**, 888-900.
- Dellaportas, P., Forster, J. and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statist. Comp.*, **12**, 27-36.
- Fernández, C., Ley, E. and Steel, M.F.J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics*, **75**, 317-343.
- George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.*, **88**, 881-889.
- George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Stat. Sinica*, **7**, 339-373.
- Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics 5, Proc. 5th Int. Meeting* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds), 609-20. Clarendon Press: Oxford, UK.
- Goswami, G. and Liu, J.S. (2007). On learning strategies for evolutionary Monte Carlo. *Statist. Comp.*, **17**, 23-38.
- Gramacy, R.B, J. Samworth, R.J. and King, R. (2007). Importance Tempering. Tech. rep. Available at: <http://arxiv.org/abs/0707.4242>
- Green, P. and Mira, A. (2001). Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, **88**, 1035-1053.
- Hans, C., Dobra, A. and West, M. (2007). Shotgun Stochastic Search for “large  $p$ ” regression. *J. Am. Statist. Assoc.*, **102**, 507-517.

- Hubner, N. *et al.* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.*, **37**, 243-253.
- Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statist. Comp.*, **11**, 313-322.
- Jasra, A., Stephens, D.A. and Holmes, C. (2007). Population-based reversible jump Markov chain Monte Carlo. *Biometrika*, **94**, 787-807.
- Liang, F., Paulo, R., Molina, G., Clyde, M.A. and Berger, J.O. (2008). Mixtures of  $g$ -priors for Bayesian variable selection. *J. Am. Statist. Assoc.*, **481**, 410-423.
- Liang, F. and Wong, W.H. (2000). Evolutionary Monte Carlo: application to  $C_p$  model sampling and change point problem. *Stat. Sinica*, **10**, 317-342.
- Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215-232.
- Natarajan, R. and McCulloch. (1998). Gibbs sampling with diffuse proper priors: a valid approach to data-driven inference?, *J. Comp. Graph. Statist.*, **7**, 267-277.
- Nott, D.J. and Green, P.J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *J. Comp. Graph. Statist.*, **13**, 141-157.
- Roberts, G.O. and Rosenthal, J.S. (2008). Example of adaptive MCMC. Tech. rep. Available at: <http://www.probability.ca/jeff/research.html>
- Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Assoc.*, **81**, 82-86.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with  $g$ -prior distributions. In *Bayesian Inference and Decision Techniques-Essays in Honour of Bruno de Finetti* (P.K. Goel and A. Zellner, eds), 233-243. Amsterdam: North-Holland.
- Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics, Proc. 1st Int. Meeting* (J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith, eds), 585-603. Valencia: University Press.

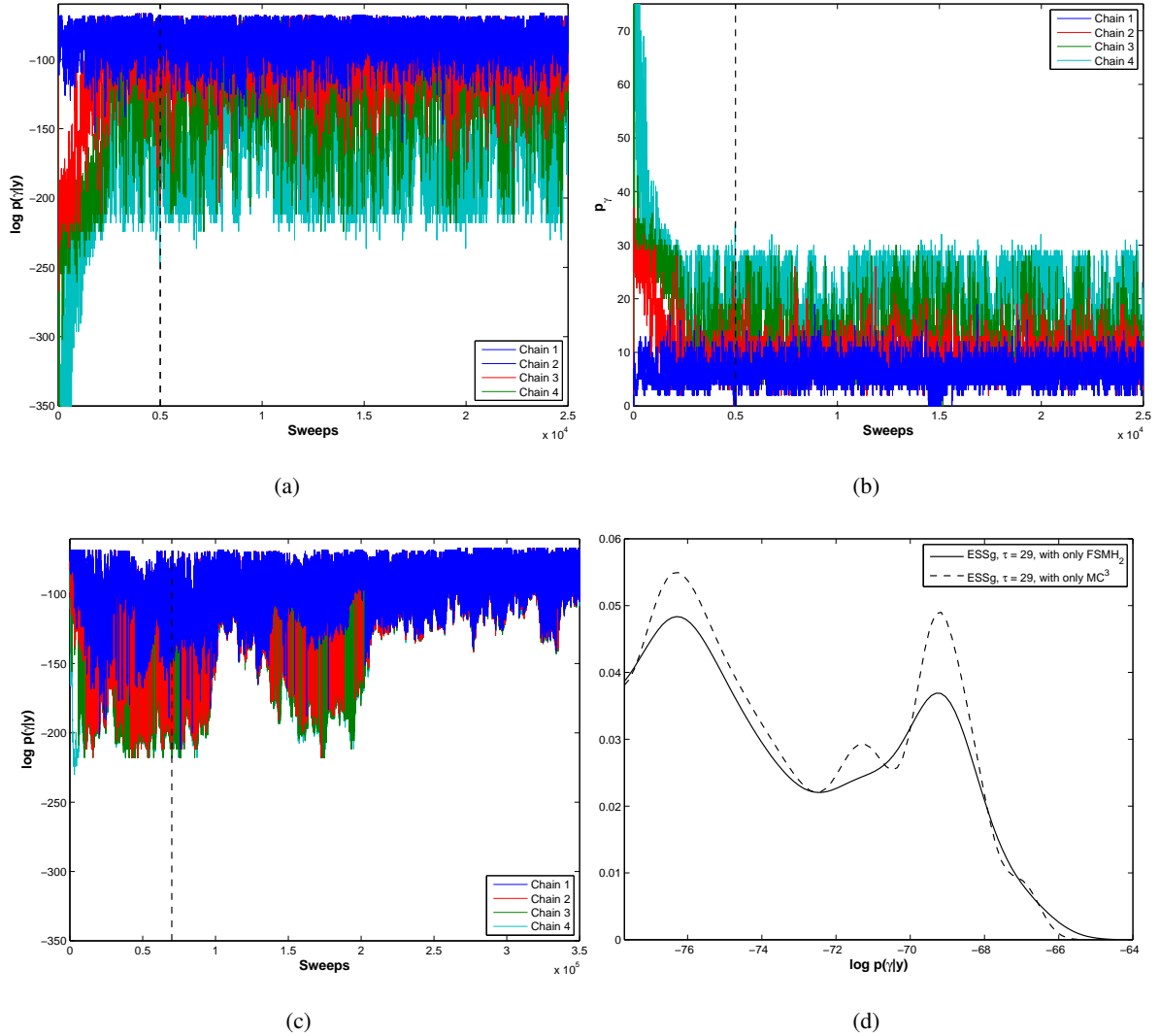


Figure 1: Top panels: (a) trace plot of the log posterior probability,  $\log p(\gamma|y)$ , and (b) model size,  $p_\gamma$ , across sweeps for the first real data example, eQTL analysis, using ESSg with  $\tau = 29$  and FSMH<sub>2</sub> as local move. Vertical dashed lines indicate the end of the burn-in. Bottom panels: (c) trace plot of the log posterior probability when MC<sub>3</sub> is used as a local move; (d) kernel densities of  $\log p(\gamma|y)$  for the retained chain in the 25 replicates of the analysis when only FSMH<sub>2</sub> and only MC<sub>3</sub> are used as a local move respectively. Plot restricted to regions of high posterior probability.

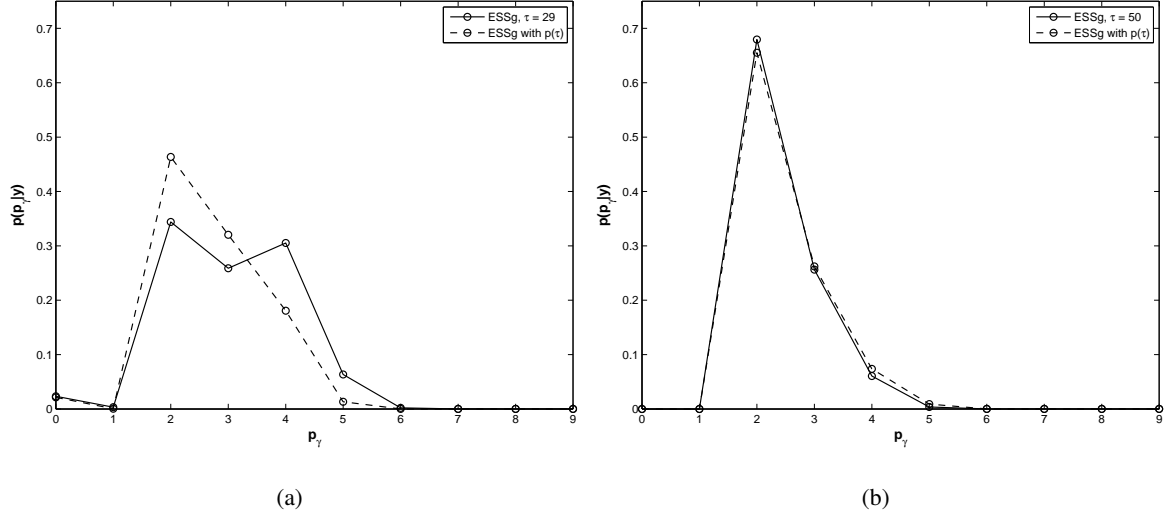


Figure 2: (a) Posterior model size for the first real data example, eQTL analysis: black solid line for ESSg with  $\tau$  fixed at 29 and black dashed line for ESSg with Z-S prior. (b) Posterior model size for mQTL analysis, second real data example, using ESSg with fixed and random  $\tau$ .

		Mode( $p_\gamma   y$ )	$E(\tau   y)$	$R_\gamma^{2*}$	$\overline{R_\gamma^{2**}}$	Stability
eQTL	ESSg, $\tau = 29$	2	—	0.716	0.704	0.257
	ESSg with $p(\tau)$	2	20.576	0.716	0.689	0.099
mQTL	ESSg, $\tau = 50$	2	—	0.843	0.843	$\approx 0$
	ESSg with $p(\tau)$	2	63.577	0.843	0.843	$\approx 0$
		Crossover	DR Exchange	ALL Exchange	Acc. rate $\tau$	Time (min.)
eQTL	ESSg, $\tau = 29$	0.214	0.534	0.671	—	28
	ESSg with $p(\tau)$	0.243	0.585	0.711	0.438	30
mQTL	ESSg, $\tau = 50$	0.214	0.514	0.669	—	302
	ESSg with $p(\tau)$	0.226	0.571	0.717	0.434	309

Table 1: Comparison between ESSg with and without the prior on  $\tau$  for the first real data example, eQTL analysis, and second example, mQTL analysis.  $R_\gamma^{2*}$  and  $\overline{R_\gamma^{2**}}$  correspond to “ $R_\gamma^2: \max p(\gamma | y)$ ” and “ $\overline{R_\gamma^2}: 1,000$  largest  $p(\gamma | y)$ ” respectively (details in Suppl. mat., Section C.3). Given the sequence of the 1,000 best  $\gamma$ s from the first chain and based on  $p(\gamma | y)$ , “Stability” is defined as the standard deviation of the corresponding  $R_\gamma^2$ s. “DR Exchange” and “ALL Exchange” stands for “delayed rejection exchange” and “all-exchange” move respectively. In the bottom part of the Table, acceptance rate for specific moves are given.