Evolutionary Stochastic Search

Leonardo Bottolo
Institute for Mathematical Sciences, Imperial College London, UK
l.bottolo@imperial.ac.uk

Sylvia Richardson*
Centre for Biostatistics, Imperial College, London, UK
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 genomics experiments. The methodology is compared with a recently proposed search algorithm in an extensive simulation study. Finally two real data examples in genomics are presented, demonstrating the performance of the algorithm in a space of up to $10,000$ covariates.

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

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 extensive 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 efficiently 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; Nott and Kohn, 2005; 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

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

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. We stress that 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 point by implementing the case of g-priors for the regression coefficients as well as the case of 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 examples with different structure on the predictors, where the number of covariates p ranges between 30 and $1,000$, and the number of samples n is both larger and smaller than p. A comparison with the recent Shotgun Stochastic Search algorithm of Hans *et al.* (2007) is presented. In Section 5 we complement the simulations results by illustrating the performance of our algorithm with two real data sets, including a challenging case where the number of predictors is extremely large ($p = 10,000$) with respect to the sample size ($n = 50$). Finally Section 6 contains some concluding remarks and a discussion of extensions.

2 Background

2.1 Variable selection

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

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

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

Suppose one wants to model the relationship between y and a subset of $x_1, \ldots, x_p$, but there is uncertainty about which subset to use. Following the usual convention of only considering models that have the intercept α, 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, with a view to gain interpretability. From a Bayesian perspective the problem is tackled by placing a constant prior density on α and a prior on β such that if $\beta_j = 0$ then the jth predictor does not appear in the expected value of y: as a result the prior structure on β depends on a latent binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)^T$, where $\gamma_j = 1$ if $\beta_j \neq 0$ and $\gamma_j = 0$ if $\beta_j = 0$. The overall number of possible models 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 γ, a Gaussian linear model can therefore be written as

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

where β_γ is the non-zero vector of coefficients extracted from β, X_γ is the design matrix of dimension $n \times p_\gamma$, $p_\gamma \equiv \gamma^T 1_p$, with columns corresponding to $\gamma_j = 1$. In the following we will assume that, apart from the intercept α, $x_1, \ldots, x_p$ 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), Berger and Molina (2005). When coupled with the latent variable γ, the conjugate prior structure of $\left(\beta_\gamma, \sigma^2\right)$ follows a normal-inverse-gamma distribution

$$p\left(\beta_\gamma \left| \sigma^2 \right.\right) = N\left(m_\gamma, \sigma^2 \Sigma_\gamma\right) \tag{2}$$

$$p\left(\sigma^2 \left| \gamma \right.\right) = p\left(\sigma^2\right) = InvGa\left(a_\sigma, b_\sigma\right) \tag{3}$$

with $a_\sigma, b_\sigma > 0$. Some guidelines how to fix the value of the hyperparameters a_σ and b_σ are provided in Cripps *et al.* (2006). The limit case of (3) when $a_\sigma \to 0$ and $b_\sigma \to 0$ corresponds to the Jeffreys' prior (e.g. Bernardo and Smith, 1994) for the error variance $p\left(\sigma^2\right) \propto \sigma^{-2}$. Taking into account both the likelihood structure (1), the prior specification for α, (2) and (3), the joint distribution of all the variables (based on further conditional independence conditions) can be written in general form as

$$p\left(y, \gamma, \alpha, \beta_\gamma, \sigma^2\right) = p\left(y \left| \gamma, \alpha, \beta_\gamma, \sigma^2 \right.\right) p(\alpha)\, p\left(\beta_\gamma \left| \gamma, \sigma^2 \right.\right) p\left(\sigma^2\right) p(\gamma). \tag{4}$$

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

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

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 with respect to positive or negative values of the coefficients (Chipman *et al.*, 2001; Clyde and George, 2004), the specification of the prior covariance Σ_γ matrix leads to at least two different classes of priors:

- When $\Sigma_\gamma = g V_\gamma$, where g is a scalar and $V_\gamma = \left(X_\gamma^T X_\gamma\right)^{-1}$, it replicates the covariance structure of the likelihood giving rise to so called g-priors first proposed by Zellner (1986).

- When $\Sigma_\gamma = c V_\gamma$, but $V_\gamma = I_{p_\gamma}$ the components of β_γ are conditionally independent and in contrast to g-priors, independent priors weaken the likelihood covariance structure.

In the following we will adopt the notation $\Sigma_\gamma = \tau V_\gamma$ as we want to cover both cases in a unified manner. Thus in the g-prior case, $\Sigma_\gamma = \tau \left(X_\gamma^T X_\gamma\right)^{-1}$ while in the independent case, $\Sigma_\gamma = \tau I_{p_\gamma}$. We will refer to τ 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), for a new proposal see Scott and Berger (2006). Here we adopt the beta-binomial prior illustrated in Kohn *et al.* (2001)

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

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_ω and b_ω can be chosen once $E(p_\gamma)$ and $V(p_\gamma)$ have been elicited. From this point of view (6) offers more flexibility than the simpler binomial model.

2.2 Priors for the variable selection coefficient τ

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 to g-priors, for independent priors the effect of $V_\gamma = I_{p_\gamma}$ on the posterior depends on the relative scale of X and standardisation of the design matrix to units of standard deviation is recommended (Chipman *et al.*, 2001). However this is not always the best procedure when the distribution X is possibly skewed, or when the columns of X are not defined on a common scale of measurement, a case which occurs often in practice when analysing joint data sets.

The second feature that makes g-priors particularly appealing is the rather simple structure of the marginal likelihood (5) with respect to the constant τ which becomes

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

where, if $m_\gamma = 0$, $S(\gamma) = ESS(\gamma) + RSS(\gamma)\,/\,(1+\tau)$ with

- $ESS(\gamma) = (y-\bar{y}_n)^T(y-\bar{y}_n) - (y-\bar{y}_n)^T X_\gamma\left(X_\gamma^T X_\gamma\right)^{-1} X_\gamma^T(y-\bar{y}_n)$, the error sum of squares.
- $RSS(\gamma) = (y-\bar{y}_n)^T X_\gamma\left(X_\gamma^T X_\gamma\right)^{-1} X_\gamma^T(y-\bar{y}_n)$, the regression sum of squares.

Given the above notation, we define $R_\gamma^2 = RSS(\gamma)\,/\,\left((y-\bar{y}_n)^T(y-\bar{y}_n)\right)$. Despite the simplicity of the marginal likelihood (7), the choice of the constant τ for g-priors is quite complex, see Fernández *et al.* (2001), George and Foster (2000), Cui and George (2008) and Liang *et al.* (2008).

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

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

with

$$p\left(\tau\right)=\frac{\left(n/2\right)^{1/2}}{\Gamma\left(1/2\right)}\tau^{-\left(1/2+1\right)}e^{-n/(2\tau)}. \tag{9}$$

The joint distribution representation (4) can now be written as

$$p\left(y,\gamma,\alpha,\beta_\gamma,\tau,\sigma^2\right)=p\left(y\left|\gamma,\alpha,\beta_\gamma,\tau,\sigma^2\right.\right)p\left(\alpha\right)p\left(\beta_\gamma\left|\gamma,\tau,\sigma^2\right.\right)p\left(\tau\right)p\left(\sigma^2\right)p\left(\gamma\right) \tag{10}$$

while the marginal likelihood has the new integral representation

$$\begin{aligned} p\left(y\left|\gamma\right.\right) &= \int p\left(y\left|\gamma,\alpha,\beta_\gamma,\tau,\sigma^2\right.\right)p\left(\alpha\right)p\left(\beta_\gamma\left|\gamma,\tau,\sigma^2\right.\right)p\left(\tau\right)p\left(\sigma^2\right)d\alpha d\beta_\gamma d\tau d\sigma^2 \\ &= \int p\left(y\left|\gamma,\tau\right.\right)p\left(\tau\right)d\tau. \end{aligned} \tag{11}$$

Liang *et al.* (2008) analyse in details Zellner-Siow priors pointing out a variety of theoretical properties. From a computational point of view, under (8) and (9), the marginal likelihood (11) is no more available in closed form which is somehow desirable in order to quickly perform a stochastic search (George and McCulloch, 1997; Chipman *et al.*, 2001). Even though in the prior set-up (9), no hyperparameters need to be specified and therefore no calibration is required, and that a Laplace approximation can be derived (Tierney and Kadane, 1986), Zellner-Siow priors never became as popular as the simpler g-prior with a constant value suitably chosen for the coefficient τ. For alternative prior specifications see also Celeux *et al.* (2006), Cui and George (2008) and Liang *et al.* (2008).

2.2.2 Independent priors

When all the variables are defined on the same scale, which often occurs in biological experiments, independent priors represent an attractive alternative to g-priors. The likelihood marginalised over α, β_γ and σ^2 has the form

$$p\left(y\left|\gamma\right.\right)\propto\tau^{-p_\gamma/2}\left|X_\gamma^TX_\gamma+\tau I_{p_\gamma}\right|^{-1/2}S\left(\gamma\right)^{-\left(2a_\sigma+n-1\right)/2}, \tag{12}$$

where $S\left(\gamma\right)=2b_\sigma+\left(y-\bar{y}_n\right)^T\left(y-\bar{y}_n\right)-\left(y-\bar{y}_n\right)^TX_\gamma\left(X_\gamma^TX_\gamma+\tau I_{p_\gamma}\right)^{-1}X_\gamma^T\left(y-\bar{y}_n\right)$. Note that (12) is computationally more demanding than (7) due to the extra determinant operator. From the above equations it is also evident that the role of the constant τ in the independent prior set-up is to regularise the quadratic form $X_\gamma^TX_\gamma$ when it is ill-conditioned.

Different approaches have been proposed to fix the value of τ. Geweke (1996) suggests to fix a different value of τ_j, $j=1,\ldots,p$, based on the idea of "substantially significant determinant" of ΔX_j with respect to Δy. Under this formulation, (12) changes accordingly with $\tau^{-p_\gamma/2}$ replaced by $\prod_{j=1}^p\tau_j^{-1/2}$ and τI_{p_γ} by T_γ which is the diagonal matrix containing the coefficients attached to the selected covariates. 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). A final approach consists in placing a prior distribution on τ (or on τ_j, $j=1,\ldots,p$) without standardising the predictors: such a strategy is illustrated for instance in Bae and Mallick (2004) and Sha *et al.* (2004).

2.2.3 Generic set-up

In order to avoid repetitions and accommodate all the cases discussed, we will describe our algorithm with a generic prior for τ, denoted by $p(\tau)$, which becomes a point mass if fixed values for τ are desired. In all the examples involving g-priors reported in Section 4.4 and 5, we will use the Zellner-Siow prior (8) with $p(\tau) = InvGam(1/2, n/2)$. For the independent prior case, we will use $\tau = 1$ or as in Section 5, a proper but diffuse prior, $p(\tau) = \mathcal{E}xp(10)$ i.e. an exponential distribution with $E(\tau) = 10$ suggested by Bae and Mallick (2004).

Finally, without loss of generality, in the following we will assume that the observed responses y have been centred with mean 0, i.e. $(y - \bar{y}_n) \equiv y$ whatever is the specification of the prior covariance matrix Σ_γ.

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. Whatever is the prior structure placed on the regression coefficients, i.e. g-priors or independent priors, since the parameters β and σ^2 are integrated out, we need to sample only the latent binary vector γ and the variable selection coefficient τ. Sampling β_γ in a forward MCMC exercise given the sampled values of γ and τ is relatively easy as illustrated in the g-prior set-up by Kohn *et al.* (2001) and in the independent prior case by Denison *et al.* (2002).

Consider the general case where a hyperprior on the variable selection coefficient τ is specified for the regression coefficients. Then the two full conditionals are

$$p(\gamma | \cdots) \propto p(y | \gamma, \tau) p(\gamma), \tag{13}$$

$$p(\tau | \cdots) \propto p(y | \gamma, \tau) p(\tau) \tag{14}$$

and sampling strategies for (13), given τ, include: Gibbs sampling (George and McCulloch, G&McC hereafter, 1993; Liu, 1996), Metropolis-Hastings (G&McC, 1997; Chipman *et al.*, 2001; Kohn *et al.*, 2001; Nott and Green, N&G hereafter, 2004; Nott and Kohn, 2005), reversible jump-type algorithms (RJ hereafter) (Denison *et al.*, 1998; Dellaportas *et al.*, 2002) and, more recently, "Shotgun Stochastic Search" (Hans *et al.*, 2007), SSS hereafter. In general the full conditional (14) is not available is closed form and a Metropolis-within-Gibbs algorithm, appears a suitable choice.

The multimodality of the model space is a known problem in variable selection (Liang and Wong, 2000; N&G, 2004; Jasra *et al.*, 2007; Hans *et al.*, 2007) and methods that tackle this problem have been proposed in the past few years: Liang and Wong (2000) suggest an extension of parallel tempering (Hukushima and Nemoto, 1996) called Evolutionary Monte Carlo, EMC hereafter; N&G (2004) deal explicitly with the problem of multicollinearity introducing a sampling scheme inspired by the Swendsen-Wang algorithm for the Ising model; Jasra *et al.* (2007) extend EMC methods to varying dimension algorithms. Finally Hans *et al.* (2007) propose a new stochastic search algorithm when $p > n$ based on the ability of SSS to explore models that are in the same neighbourhood in order to quickly find the best combination of predictors.

Given the full conditionals (13) and (14), we show that the two issues, namely the multimodality of model space and the dependence between γ and τ, can be solved jointly by applying some suitable "tempering strategies" directly on $p\left(y\left|\gamma, \tau\right.\right)$.

In the following Subsections we describe an efficient MCMC algorithm to sample from the two full conditionals. In particular we apply an EMC scheme to (13), giving rise to a multiple chains scheme, and an adaptive Metropolis-within-Gibbs sampler to (14). This latter strategy automatically adapts the variance of the proposal distribution to the model space visited by the algorithm, avoiding the time consuming tuning of the proposal in a pre-run MCMC. Moreover it overcomes the known problem of the Gibbs sampler when the prior is proper, but diffuse (Natarajan and McCulloch, 1998). Further discussions of our motivation for sampling τ are given in more details in Subsection 3.2.

3.1 EMC sampler for $p\left(\gamma\left|\cdots\right.\right)$

A large amount of literature exists on EMC and notable examples are: Liang and Wong (2000, 2001), Liu (2001), Jasra *et al.* (2007), Goswami and Liu (2007) amongst others. The basic idea of EMC is that it encompasses the positive features of simulated annealing and genetic algorithm inside a MCMC scheme. It is characterised by a population of L Markov Chains that are simulated in parallel, each of which attached with a different temperature. At each EMC sweep, the population of chains is updated using a variety of moves: "local moves" based on the mutation operator, the ordinary Metropolis-Hastings or Gibbs update on every chain $l = 1, \ldots, L$; and "global moves" that include selection of the chains to swap based on some probabilistic measures of distance between them, crossover operator, i.e. partial swap of the current state between different chains, and exchange operator, full state swap between chains. Global moves are crucial because they allow the algorithm to jump from one local mode to another.

The logarithm transformation of $p\left(\gamma\left|\cdots\right.\right)$, $f\left(\gamma\left|\tau\right.\right) = \log p\left(y\left|\gamma, \tau\right.\right) + \log p\left(\gamma\right)$ is the conditional target function. The population corresponds to a set of chains, that are retained simultaneously. We will use the double indexing $\gamma_{l,j}$, $l = 1, \ldots, L$ and $j = 1, \ldots, p$ to denote the jth latent binary indicator in the lth chain. Moreover we indicate by $\gamma_l = (\gamma_{l,1}, \ldots, \gamma_{l,p})$ the vector of binary indicators that characterise the state of the lth member of the population.

In what follows, we will only sketch the rationale behind all the moves that we found useful to implement. For the "large p, small n" paradigm and complex predictor spaces, we believe that using a wide portfolio of moves is needed and offers sure guarantee of mixing.

Local moves: mutation operator and Fast Scan Metropolis Hastings sampler

Given τ we implemented the simple MC^3 idea of Madigan and York (1995), also used by Brown *et al.* (1998, 2002) where add/delete and swap moves are used to update the latent binary vector γ_l. However, as noted in Hans *et al.* (2007) when p is large relatively to p_γ, the algorithm spends most of the time trying to add rather than delete a variable: given the lth chain with p_{γ_l} the size of the current model, the probability of selecting a variable to be deleted, is p_{γ_l}/p and if p is large with respect to p_γ the algorithm will spend a very large amount of time trying to add a variable before selecting a variable to be deleted.

On the other hand, Gibbs sampling or Metropolised Gibbs sampling (Liu, 1996), are not affected by this problem since the state of the lth chain is updated by sampling from

$$p\left(\gamma_{l,j}=1\left|y,\gamma_{l,j^-},\tau\right.\right)^{1/t_l}\propto\exp\left\{\left(\log p\left(y\left|\gamma_{l,j}^{(1)},\tau\right.\right)+\log p\left(\gamma_{l,j}=1\left|\gamma_{l,j^-}\right.\right)\right)/t_l\right\}, \tag{15}$$

where t_l is the temperature attached to the lth chain, $1=t_1<t_2<\cdots<t_L$, γ_{l,j^-} indicates for the lth chain all the variables, but the jth, $j=1,\ldots,p$ and $\gamma_{l,j}^{(1)}=\left(\gamma_{l,1},\ldots,\gamma_{l,j-1},\gamma_{l,j}=1,\gamma_{l,j+1},\ldots,\gamma_{l,p}\right)^T$. Now the main problem related to Gibbs sampling is the large number of models it evaluates if a full Gibbs cycle through $j=1,\ldots,p$ 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 $S\left(\gamma\right)$, equation (7) or (12), making practically impossible to apply 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 models because if p_γ is much smaller than p and given $\gamma_j=0$, it is likely that γ_j "regenerates" as 0.

In the case where p is large, we thus consider instead of the standard MC^3 add/delete, swap moves, two novel Fast Scan Metropolis-Hastings schemes, FSMH hereafter, specialised for EMC/parallel tempering. They are computationally less demanding than a full Gibbs sampling on all γ_j and do not suffer from the problem highlighted before for MC^3 and RJ-type algorithms when p is large with respect to p_γ. The idea behind the FSMH is to use an acceptance/rejection step (which is very fast to evaluate) to choose the indices where to perform the time consuming evaluation of the Gibbs-like step. One key point of our FSMH sampler is that the probability used in the acceptance/rejection step is "adaptive" and based not only on the current chain model size p_{γ_l}, but also on the temperature t_l attached to the lth chain. Full details of the two FSMH schemes that we are using are given in the Appendix.

Global move: crossover operator

The first step of this move consists in selecting the pair of chains to be operated on. We compute a probability equal to the weights of the "Boltzmann probability"

$$p_t\left(\gamma_l\left|\tau\right.\right)=\frac{\exp\left\{f\left(\gamma_l\left|\tau\right.\right)/t\right\}}{F_t}, \tag{16}$$

where $F_t=\sum_{l=1}^{L}\exp\left\{f\left(\gamma_l\left|\tau\right.\right)/t\right\}$ and rank all the chains according to this. The first chain is chosen randomly with normalised Boltzmann weights (16) and the second one is chosen randomly from the top (usually) half (rounded up) of the chains (excluding the first one). We refer to this first step as "selection operator".

Suppose that two "offsprings" are then generated from the parental chains according to some crossover operator described below. The new proposed population of chains $\boldsymbol{\gamma}'=\left(\gamma_1,\ldots,\gamma_l',\ldots,\gamma_r',\ldots,\gamma_L\right)$ is accepted with probability

$$\alpha\left(\boldsymbol{\gamma}\to\boldsymbol{\gamma}'\right)=\min\left\{1,\frac{\exp\left\{f\left(\gamma_l'\left|\tau\right.\right)/t_l+f\left(\gamma_r'\left|\tau\right.\right)/t_r\right\}}{\exp\left\{f\left(\gamma_l\left|\tau\right.\right)/t_l+f\left(\gamma_r\left|\tau\right.\right)/t_r\right\}}\frac{Q_t\left(\boldsymbol{\gamma}'\to\boldsymbol{\gamma}\left|\tau\right.\right)}{Q_t\left(\boldsymbol{\gamma}\to\boldsymbol{\gamma}'\left|\tau\right.\right)}\right\}, \tag{17}$$

where $Q_t\left(\boldsymbol{\gamma}\to\boldsymbol{\gamma}'\left|\tau\right.\right)$ is the proposal probability.

In the following we will assume that four different crossover operators are selected at random at every EMC sweep: 1-point crossover, uniform crossover and 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 offsprings' states 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: the idea behind is that if variables are correlated their binary latent values should be similar and trying a 1-point crossover which introduces a random crossover point can destroy the correlation structure. Therefore the block crossover can be seen as a k-points crossover whose crossover points are not random but defined from the correlation structure of the covariates. The level of pairwise correlation above which variables are considered part of the same group is arbitrary but we fixed it at 0.80.

Global move: exchange operator

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 the second proposed chain receives the whole first state chain $\gamma_r' = \gamma_l$, respectively.

In order to achieve a good acceptance rate, 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 one on Gibbs distribution over all possible chains pairs (Calvo, 2005). Both of them perform well in the simulated examples, see Subsections 4.2 and 4.4 and real data applications, see Section 5.

1. The delayed rejection exchange operator tries first to swap the state of the chains that are usually far apart in the temperature ladder, but, once the proposed move has been rejected, it performs a more traditional (uniform) adjacent pair selection, increasing the overall mixing between chains on one hand without drastically reducing the acceptance rate on the other. However its flexibility comes at some extra computational costs and in particular the additional evaluation of the pseudo move necessary to maintain detailed balance (Green and Mira, 2001).

2. Alternatively, we attempt a bolder "all-exchange" operator. Swapping the state of two chains that are far apart in the temperature ladder speeds up the convergence of the simulation since it replaces several adjacent swaps with a single move. However, this move can be seen as a rare event whose acceptance probability is low and unknown. Since the full set of possible exchange moves is finite and discrete, it is easy and computationally inexpensive to calculate all the $L\left(L-1\right)/2$ exchange acceptance rates between all chains' pairs, inclusive the rare ones, $\tilde{p}_{l,r} = \exp\left\{f\left(\gamma_l\left|\tau\right.\right)/t_l - f\left(\gamma_r\left|\tau\right.\right)/t_r\right\}$, $(l, r < l)$. To maintain detailed balance condition, the possibility not to perform any exchange (rejection) must be added with unnormalised probability one. Finally the chains whose states are swopped are selected at random with probability equal to

$$p_h = \frac{\tilde{p}_h}{\sum_{h=1}^{1+L(L-1)/2} \tilde{p}_h}, \tag{18}$$

where in (18) each pair $(l, r < l)$ is denoted by a single number h, $\tilde{p}_h = \tilde{p}_{l,r}$, including the rejection move, $h = 1$.

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 based on the complexity of the problem (Liang and Wong, 2001): in particular we choose $L = \min\{5, E(p_\gamma)\}$. Our motivation is to relate the number of models the algorithm simultaneously evaluates at every EMC sweep to the expected model size. Secondly, we fix a first stage temperature ladder according to a geometric scale such that $t_{l+1}/t_l = b$, $b > 1$, $l = 1, \ldots, L$ with b relatively large, for instance $b = 4$. Finally, we adopt a strategy similar to the one described in Roberts and Rosenthal (2007), but restricted to the burn-in stage, monitoring only the acceptance rate of the delayed rejection exchange operator. After the kth "batch" of EMC sweeps, to be chosen but usually set equal to 100, we update b_k, the value of the constant b up to the kth batch, by adding or subtracting an amount δ_b such that the acceptance rate of the delayed rejection exchange operator is as closed as possible to 0.50 (Liu, 2001; Jasra *et al.*, 2007), $b_{k+1} = 2^{\log_2 b_k \pm \delta_b}$. Specifically the value of δ_b is chosen such that at the end of the burn-in period the value of b can be 1, i.e. all the chains has the same baseline temperature. To be precise, we fix the value of δ_b as $\log_2(b_1)/\tilde{K}$, where b_1 is the first value assigned to the geometric ratio and $\tilde{K}$ is the number of batches in the burn-in period. We stress that this is not an adaptive EMC scheme since we just adopt the iterative procedure to reach the desired acceptance rate during the burn-in although a complete adaptive EMC scheme would not difficult to implement.

3.2 Adaptive Metropolis-within-Gibbs for $p(\tau \mid \cdots)$

Various sampling strategies can be used to sample from the posterior distribution of the variable selection coefficient τ. However, whatever is the prior specification of the prior covariance matrix Σ_γ, the random variable τ can be seen as a nuisance parameter. The easiest way to integrate it out is through a Laplace approximation (Berger *et al.*, 1999) or using a numerical integration such as a quadrature on an infinite interval.

We do not pursue these strategies and the reasons can be summarised as follows. Integrating out τ in the population implicitly assumes that every chain has its own value of the latent binary vector γ_l and variable selection coefficient τ_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 γ_l is conditioned to the chain specific value $\hat{\tau}_{\gamma_l}$. For example, considering for the g-prior case the hyper-g prior proposed by Liang *et al.* (2008), it is easy to show that

$$\hat{\tau}_{\gamma_l} = \max\left\{ \frac{R^2_{\gamma_l} / (p_{\gamma_l} + a_\tau)}{\left[2b_\sigma/(y^T y) + \left(1 - R^2_{\gamma_l}\right)\right] / \left[2a_\sigma + n - 1 - (p_{\gamma_l} + a_\tau)\right]} - 1, 0 \right\}$$

with $a_\tau > 2$. Now, since the chains with higher temperature are allowed to freely explore the model space, it is likely that most of the times $R^2_{\gamma_l} \approx 0$ which implies that also $\hat{\tau}_{\gamma_l} \approx 0$: from (7) it is evident that when the variable selection coefficient is very small, the marginal likelihood depends weakly on X_{γ_l}. In this situation chains attached to high temperatures will experience a very unstable behaviour, making the convergence in the product space hard to reach. The same situation arises also with the Zellner-Siow prior (8). The second problem is a direct consequence of the former: since chains at high temperature are unstable, global EMC type moves are difficult to implement, reducing the overall acceptance rate of both crossover and exchange operators. In addition, if an automatic tuning of temperature ladder is applied, chains will increasingly be placed at a closer distance in the temperature ladder in order to compensate the low acceptance rate of the global operators, negating the purpose of the EMC scheme.

In this paper the convergence is reached instead in the product space $\prod_{l=1}^{L} p\left(\gamma_l, \tau_1 \left| y, X\right.\right)$ i.e. the whole population is conditioned to the value of $\tau = \tau_1$ sampled from the first chain. This strategy will alleviate the problems highlighted before allowing a faster convergence and a better mixing among the chains. The procedure just described comes with an extra cost i.e. sampling the value of τ. However this step is inexpensive relatively to the cost required to sample γ_l, $l = 1, \ldots, L$. There are several strategies that can be used to sample τ from (14). We found useful to apply the idea of adaptive Metropolis-within-Gibbs described in Roberts and Rosenthal (2007). In our set-up it has several benefits; amongst others it avoids the known problems faced by the Gibbs sampler when the prior is proper, but relative flat as it can happen for the Zellner-Siow prior when n is large or for the independent case considered by Bae and Mallick (2004).

In the following we provide some details of the implementation. Since τ is defined on the real positive axis we propose the new value of τ on the logarithm scale. In particular we use as proposal the normal distribution centred at the current value of $\log\left(\tau\right)$ in the g-prior case and 0 in the independent case. The variance of the proposal distribution is controlled as illustrated in Roberts and Rosenthal (2007): every 100 EMC sweeps, the same value of EMC sweeps used in the temperature placement, we monitor the acceptance rate of the Metropolis-within-Gibbs algorithm: if it is lower than the optimal acceptance rate, i.e. 0.44, a constant $\delta_\tau(K)$ is added to ls_k, the log standard deviation of the proposal distribution in the kth batch of EMC sweeps. The value of the constant to be added or subtracted is rather arbitrary but we found useful to fix it as $\left|ls_1 - 5\right| / \tilde{K}$ i.e. during the burn-in the log standard deviation should be able to reach any values at a distance ±5 in log scale from the initial value of ls_1 usually set equal to zero. Finally the diminishing adaptation condition is obtained imposing $\delta_\tau\left(K\right) = \min\{\left|ls_1 - 5\right| / \tilde{K}, K^{-1/2}\}$, where $\tilde{K}$ and K are the number of batches in the burn-in and in the whole EMC scheme respectively. The bounded conditions are not a problem since the sequence of the standard deviations of the proposal distribution stabilises almost immediately, see Subsection 4.4 and Section 5, but we fix them equal to $M_1 = -10$ and $M_2 = 10$ such that $ls_k \in [M_1, M_2]$.

3.3 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 ESSg while we use ESSi if independent priors are selected. Moreover the same notation is used for cases where τ is fixed or is given a prior distribution. We also assume that the response vector and the design matrix have both been centred. Based on the two full conditionals (13) and (14) and the local and global operators introduced earlier, our ESS can be summarised as follows.

- Given τ, sample the population's states $\boldsymbol{\gamma}$ from the two steps:

  (i) With probability $\pi = 0.5$ perform local move and with probability $1 - \pi$ apply at random one of the four crossover operators: 1-point, uniform, block and adaptive crossover. If local move is selected, apply MC^3 if $n > p$ and if $p \geq n$ use FSMH sampling scheme 2 (see Appendix) independently for each chain. 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 (a) or the all-exchange operator (b) with equal probability. During the burn-in, only select the delayed rejection exchange operator.

- When τ is not fixed but has a prior $p(\tau)$, given the latent binary vector $\gamma = \gamma_1$ (first chain), sample τ from an adaptive Metropolis-within-Gibbs sampling (Section 3.2).

From a computational point of view, we used the same fast form for update $S(\gamma)$ as Brown *et al.* (1998, 2002), based on QR decomposition of $ESS(\gamma)$. Besides its numerical benefits, the QR decomposition automatically deals with the case $p_\gamma \geq n$, a situation that might occur for some chains at high temperature.

4 Simulation study

In this Section we illustrate the performance of ESS described in Subsection 3.3 in a variety of simulated examples. We firstly analyse the simulated examples with ESSi the version of our algorithm which assumes independent priors, $\Sigma_\gamma = \tau I_{p_\gamma}$, so as to enable comparisons with SSS which also implements an independent prior. Moreover, in order to make to comparison with SSS fair, in the simulation study only the first step of the algorithm described in Subsection 3.3 is performed, with τ fixed at 1. As in SSS, standardisation of the covariates is done before running ESSi. We run ESSi and SSS 2.0 (Hans *et al.*, 2007) for the same number of sweeps (22,000) and with matching hyperparameters on the model size.

Secondly, to discuss the mixing properties of ESS when a prior $p(\tau)$ is defined on τ, we implement both the g-prior and independent prior set-up for a particular simulated experiment. To be precise in the former case we will use the Zellner-Siow prior (8) and for the latter we will specify a proper but diffuse exponential distribution as suggested by Bae and Mallick (2004).

4.1 Simulated experiments

We apply ESS with independent priors to an extensive and challenging range of simulated examples with τ fixed at 1: the first three examples (Ex1-Ex3) consider the case $n > p$ while the remaining three (Ex4-Ex6) have $p > n$. Moreover in all examples, except the last one, we simulate the design matrix, creating more and more intricated correlation structures between the covariates in order to test the proposed algorithm in different and increasingly more realistic scenarios. In the last example, we use, as design matrix, a genetic region spanning 500-kb from the HapMap project (Altshuler *et al.*, 2005).

Simulated experiments Ex1-Ex5 share in common the way we build X. In order to create moderate to strong correlation, we found useful referring to the second simulated example in G&McC (1993) and in G&McC (1997): throughout we call X_1 ($n \times 60$) and X_2 ($n \times 15$) the design matrix obtained from these two examples. Then, as in N&G (2004) Example 2, more complex structures are created by placing side by side combinations of X_1

and/or X_2, with different sample size. We will vary the number of samples n in X_1 and X_2 as we construct our examples. The levels of β are taken from the simulation study of Fernández *et al.* (2001), while the number of true effects, p_γ, with the exception of Ex3, varies from 5 to 16. Finally the simulated error variance ranges from 0.05^2 to 2.5^2 in order to vary the level of difficulty for the search algorithm. Throughout we only list the non-zero β_γ and assume that $\beta_{\gamma^-} = 0^T$. The six examples can be summarised as follows:

Ex1: $X = X_1$ is a matrix of dimension 120×60, where the responses are simulated from (1) using $\alpha = 0$, $\gamma = (21, 37, 46, 53, 54)^T$, $\beta_\gamma = (2.5, 0.5, -1, 1.5, 0.5)^T$, and $\varepsilon \sim N\left(0, 2^2 I_{120}\right)$. In the following we will not refer to the intercept α any more since, as described in Subsection 2.1, we consider y centred and hence there is no difference in the results if the intercept is simulated or not. This is the simplest of our example, although, as reported in G&McC (1993) the average pairwise correlation is about 0.5, making it already hard to analyse by standard stepwise methods.

Ex2: This example is taken directly from N&G (2004), Example 2, who first introduce the idea of combining simpler "building blocks" to create a new matrix X : in their example $X = \left[X_2^{(1)} X_2^{(2)}\right]$ is a 300×30 matrix, where $X_2^{(1)}$ and $X_2^{(2)}$ are of dimension 300×15 and have each the same structure as X_2. Moreover $\gamma = (1, 3, 5, 7, 8, 11, 12, 13)^T$, $\beta_\gamma = (1.5, 1.5, 1.5, 1.5, -1.5, 1.5, 1.5, 1.5)^T$ and $\varepsilon \sim N\left(0, 2.5^2 I_{300}\right)$. We chose this example for two reasons: firstly, since the correlation structure in X_2 is very involved, we test the proposed algorithm under strong and complicated correlations between the covariates; secondly, since y is not simulated from the second "block", we are interested to see if the proposed algorithm does *not* select any variable that belongs to the second group.

Ex3: As in G&McC (1993), Example 2, $X = X_1$, is a 120×60 matrix, $\beta = (\beta_1, \ldots, \beta_{60})^T$, $(\beta_1, \ldots, \beta_{15}) = (0, \ldots, 0)$, $(\beta_{16}, \ldots, \beta_{30}) = (1, \ldots, 1)$, $(\beta_{31}, \ldots, \beta_{45}) = (2, \ldots, 2)$, $(\beta_{46}, \ldots, \beta_{60}) = (3, \ldots, 3)$ and $\varepsilon \sim N\left(0, 2^2 I_{120}\right)$. The motivation behind this example is to test the strength of the proposed algorithm to select a subset of variables which is large with respect to p while preserving the ability *not* to choose any of the first 15 variables.

Ex4: The design matrix X, 120×300, is constructed as follows: firstly we create a new 120×60 "building block", X_3, combining X_2 and a smaller version of X_1, X_1^*, a 120×45 matrix simulated as X_1, such that $X_3 = [X_2 X_1^*]$. Secondly we place side by side five copies of X_3, $X = \left[X_3^{(1)} X_3^{(2)} X_3^{(3)} X_3^{(4)} X_3^{(5)}\right]$: the new design matrix alternates blocks of covariates of high and complicated correlation, as in G&McC (1997), with regions where the correlation is moderate as in G&McC (1993). We simulate the response selecting 16 variables from X, $\gamma = (1, 11, 30, 45, 61, 71, 90, 105, 121, 131, 150, 165, 181, 191, 210, 225)^T$ such that every pair belongs alternatively to X_2 or X_1. We simulate y using $\beta_\gamma = (2, -1, 1.5, 1, 0.5, 2, -1, 1.5, 1, 0.5, 2, -1, -1, 1.5, 1, 0.5)^T$ with $\varepsilon \sim N\left(0, 2.5^2 I_{120}\right)$. This example is challenging in view of the correlation structure, the number of covariates $p > n$ and the different levels of the effects.

Ex5: This is the most challenging example that we simulated and it is based on the idea of contaminated models. The matrix X, 200×1000, is $X = \left[X_3^{(1)} X_3^{(2)} X_3^{(3)} X_1^{**} X_3^{(4)} X_3^{(5)} X_3^{(6)} X_3^{(7)} X_3^{(8)}\right]$, with X_1^{**}, a 200×520 larger version of X_1. We partitioned the responses such that $y = [y_1 y_2]^T$: y_1 is simulated from "model 1"

($\gamma^1 = (701, 730, 745, 763, 790, 805, 825, 850, 865, 887)$ and $\beta^1_\gamma = (2, -1, 1.5, 1, 0.5, 2, -1, 1.5, 2, -1)$) while y_2 is simulated from "model 2" ($\gamma^2 = (1, 38, 63, 98, 125)$ and $\beta^2_\gamma = (2, -1, 1.5, 1, 0.5)$). Finally, fixing $\varepsilon \sim N\left(0, 0.05^2 I_{200}\right)$ and the sample size in the two models such that y_1 and y_2 are vectors of dimension 1×160 and 1×40 respectively, y is retained if, given the sampling variability, we find $R^2_{\gamma^1} \geq 0.6$ and $R^2_{\gamma^1}/8 \leq R^2_{\gamma^2} \leq R^2_{\gamma^1}/10$: in this way we know that "model 1" accounts for most of the variability of y, but without a negligible effect for "model 2". In this example, we measure the ability of the proposed algorithm to recognise the most promising model and therefore being robust to contaminations. However since ESS can easily jump between local modes we are also interested to see if "model 2" is selected.

Ex6: The last simulated example is based on phased genotype data from HapMap project (Altshuler *et al.*, 2005), region ENm014, Yoruba population: the data set originally contained 1,218 SNPs (Single Nucleotide Polymorphism) for 120 chromosomes, but after eliminating redundant variables, the design matrix reduced to 120×775. While in the previous examples a "block structure" of correlated variables is artificially constructed, in this example blocks of linkage disequilibrium (LD) derive naturally from genetic forces, with a slow decay of the level of pairwise correlation between SNPs. Finally we chose $\gamma = (50, 75, 140, 200, 300, 400, 500, 650, 700, 770)^T$ such that the effects are visually inside blocks of LD, with their size simulated from $\beta_\gamma \sim N\left(0, 3^2 I_{10}\right)$ with $\varepsilon \sim N\left(0, 0.10^2 I_{120}\right)$. Since the simulated effects can range roughly between $(-6, 6)$, this will allow us to test also the ability of ESSi to select small effects.

We conclude this Subsection by reporting how we conducted the simulation experiment: every example from Ex1 to Ex6 has been replicated 25 times and the results presented for example Ex1 to Ex5 are averaged over the 25 replicates. For Ex6 the effects size change so average across replicated is only done for the mixing properties. ESSi with τ =1 was applied to each example/sample, recording the visited sequence of γ_1 for $20,000$ sweeps after a burn-in of $2,000$ required for the automatic tuning of the temperature placement, Subsection 3.1. With the exception of Ex2 and Ex3, where we used an indifferent prior, $p(\gamma) = (1/2)^p$, we analysed the remaining examples setting $E(p_\gamma) = 5$ with $V(p_\gamma) = E(p_\gamma)(1 - E(p_\gamma)/p)$ which corresponds to a binomial prior over p_γ. In order to establish the sensitivity of the proposed algorithm to the choice of $E(p_\gamma)$ we also analysed Ex1 fixing $E(p_\gamma) = 10$ and 20. Moreover in all the examples we chose $L = 5$ in keeping with the expected model size, with the starting value of γ chosen at random. The remaining two hyperparameters to be fixed, namely a_σ and b_σ, are set equal to $a_\sigma = 10^{-6}$ and $b_\sigma = 10^{-3}$ as in Kohn *et al.* (2001) which corresponds to a relative uninformative prior.

4.2 Mixing properties of ESSi

In this Subsection we report some stylised facts about the performance of the ESSi with τ fixed at 1. Figure 1, top panels, shows for one of the replicates of Ex1, the overall mixing properties of ESSi. As expected, the chains attached to higher temperatures shows more variability. Albeit the convergence is reached in the product space $\prod_{l=1}^{L} p(\gamma_l | y, X)$, by visual inspection 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 their log posterior probabilities overlap nicely, allowing effective exchange of information between the chains.

Figure 1, bottom panels, shows the trace plot of the log posterior and the posterior model size for a replicate of Ex4. We can see that also in the case $p > n$, the chains mix and overlap well with no gaps between them, the automatic tuning of the temperature ladder being able to improve drastically the performance of the algorithm.

This effective exchange of information is demonstrated in Table 1 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 rate of the (delayed rejection) exchange operator which stays very stable and close to the target: for instance in Ex4 ($p = 300$) and Ex6 ($p = 775$) the mean and standard deviation of the acceptance rate are 0.517 (0.105) and 0.497 (0.072) while in Ex5 ($p = 1,000$) we have 0.505 (0.013): the higher variability in Ex4 being related to the model size p_γ.

With regards to the crossover operators, again we observe stability across all the examples. Moreover, in contrast to Jasra *et al.* (2007), when $p > n$, the crossover average acceptance rate across the 5 chains is quite stable between 0.147, Ex4, and 0.193, Ex6 (with the lower value in Ex4 here again due to p_γ): within our limited experiments, we believe that the good performance of crossover operator is related to the selection operator, see Subsection 3.1.

Some finer tuning of the temperature ladder could still be performed as there seems to be an indication that fewer global moves are accepted with the higher temperature chain, see Table 2, where swapping probabilities for each chain are indicated. Note that the observed frequency of successful swaps is not far from the case where adjacent chains are selected to swap at random with equal probability. Other measures of overlapping between chains (Liang and Wong, 2000; Iba 2001), based on a suitable index of variation of the target function $f(\gamma) = \log p(y|\gamma) + \log p(\gamma)$ across sweeps, confirm the good performance of ESSi. Again some instability is present in the high temperature chains, see in Table 2 the overlapping index between chains $3, 4$ and $4, 5$ in Example 3 to 6.

In order to overcome this problem, we also tried a different temperature placement approach based exclusively on the overlapping index between chains suggested by Iba (2001). However, while this strategy can give a better mixing between chains, it is difficult to implement in a fully automatic way: changing the temperature attached to one chain in order to reach the desired overlapping between two consecutive chains, implicitly modifies also the overlapping index among all the remaining chains: in the population the overlapping is controlled by $L - 1$ depending temperatures which are difficult to set together. Our temperature placement implementation instead depends on just one parameter, the geometric ratio b which is easy to handle.

In Ex1, we also investigate the influence of different values of the prior mean of the model size. We found that the average (standard deviation in brackets) acceptance rate across replicates for the (delayed rejection) exchange operator ranges from 0.493 (0.043) to 0.500 (0.040) for different values of the prior mean on the model size, while the acceptance rate for the crossover operator ranges from 0.249 (0.021) to 0.271 (0.036). This strong stability is not surprising because the automatic tuning modifies the temperature ladder in order to compensate for $E(p_\gamma)$. Finally we notice that the acceptance rates for the local move, when $n > p$, increases with higher values of the prior mean model size, showing that locally the algorithm moves more freely with $E(p_\gamma) = 20$ than with $E(p_\gamma) = 5$.

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

4.3 Performance of ESSi and comparison with SSS

Performance of ESSi

We conclude this Section by discussing in details the overall performance of ESSi with respect to the selection of the true simulated effects. As a first measure of performance, we report for all the simulated examples the marginal posterior probability of inclusion as described in G&McC (1997) and Hans *et al.* (2007). In the following, for ease of notation, we drop the chain subscript index and we exclusively refer to the first chain. To be precise, we evaluate the marginal posterior probability of inclusion as:

$$p\left(\gamma_j = 1\left|y\right.\right) \simeq C^{-1} \sum_{k=1,\ldots,K} 1_{\left(\gamma_j^{(k)}=1\right)}\left(\gamma\right) p\left(y\left|\gamma^{(k)}\right.\right) p\left(\gamma^{(k)}\right) \tag{19}$$

with $C = \sum_{k=1,\ldots,K} p\left(y\left|\gamma^{(k)}\right.\right) p\left(\gamma^{(k)}\right)$ and K the number of sweeps after the burn-in. The posterior model size is similarly defined, $p\left(p_\gamma\left|y\right.\right) \simeq C^{-1} \sum_{k=1,\ldots,K} 1_{\left(\left|\gamma^{(k)}\right|=p_\gamma\right)}\left(\gamma\right) p\left(y\left|\gamma^{(k)}\right.\right) p\left(\gamma^{(k)}\right)$, with C as before. Besides plotting the marginal posterior inclusion probability (19) averaged across sweeps and replicates for our simulated examples, we will also compute the interquatile range of (19) across replicates as a measure of variability.

In order to thoroughly compare the proposed ESS algorithm to SSS (Hans *et al.*, 2007), we present also some other measures of performance based on $p\left(\gamma\left|y\right.\right)$ and R_γ^2 : first we rank $p\left(\gamma\left|y\right.\right)$ in decreasing order and record the indicator γ that corresponds to the maximum and $1,000$ largest $p\left(\gamma\left|y\right.\right)$ (after burn-in). Given the above set of latent binary vectors, we then compute the corresponding R_γ^2 leading to "R_γ^2: $\max p\left(\gamma\left|y\right.\right)$" as well as the mean R_γ^2 over the $1,000$ largest $p\left(\gamma\left|y\right.\right)$, "$\overline{R_\gamma^2}$: $1,000$ largest $p\left(\gamma\left|y\right.\right)$", both quantities averaged across replicates. Moreover the actual ability of the algorithm to reach (quickly) regions of high posterior probability and persist on them is monitored: given the sequence of the $1,000$ best γ (based on $p\left(\gamma\left|y\right.\right)$), the standard deviation of the corresponding R_γ^2s shows how stable is the searching strategy at least for the top ranked (not unique) posterior probabilities: averaging over the replicates, it provides an heuristic measures of "stability" of the algorithm. Finally we report the average computational time (in minutes) across replicates of ESSi written in Matlab code and run on a 2MHz CPU with 1.5 Gb RAM desktop computer and of SSS version 2.0 on the same computer.

Comparison with SSS

Figure 2 presents the marginal posterior probability of inclusion for ESSi with $\tau = 1$ averaged across replicates and, as a measure of variability, the interquantile range, blue left triangles and vertical blue solid line respectively. In general the covariates with non zero effects have high marginal posterior probability of inclusion in all the examples: for example in Ex3, Figure 2 (a), the proposed ESSi algorithm, blue left triangle, is able to perfectly select the last 45 covariates, while the first 15, which do not contribute to y, receive small marginal posterior probability. It is interesting to note that this group of covariates, $(\beta_1, \ldots, \beta_{15}) = (0, \ldots, 0)$, although correctly recognised having no influence on y, show some variability across replicates, vertical blue solid line: however, this is not surprising since independent priors are less suitable in situations where all the covariates are mildly-strongly correlated as in this simulated example. On the other hand the second set of covariates with small effects, $(\beta_{16}, \ldots, \beta_{30}) = (1, \ldots, 1)$, are univocally detected. The ability of ESSi to select variables with small effects is also evident in Ex6, Figure 2 (d), where the two smallest coefficients, $\beta_2 = 0.112$ and $\beta_{10} = 0.950$ (the second and last respectively from left to right), receive from high to very high marginal posterior probability (and similarly for the other replicates, data not shown). In some cases however, some covariates attached with small effects are missed (e.g. Ex4, Figure 2 (b), the last simulated effect which is also the smallest, $\beta_{16} = 0.5$, is not detected). In this situation however the vertical blue solid line indicates that for some replicates, ESSi is able to assign small values of the marginal posterior probability giving evidence that ESSi fully explore the whole space of models.

Superimposed on all pictures of Figure 2 are the median and interquantile range across replicates of $p\left(\gamma_j = 1\,|y\right)$, $j = 1, \ldots, p$, for SSS, red right triangles and vertical red dashed line respectively. We see that there is good agreement between the two algorithms in general, with in addition evidence that ESSi 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. For instance in Ex3, where the last 30 covariates accounts for most of R^2_γ, SSS has difficulty to detect $(\beta_{16}, \ldots, \beta_{30})$, while in Ex6, it misses $\beta_2 = 0.112$, the smallest effect and surprisingly also $\beta_4 = -2.595$ assigning a very small marginal posterior probability (and in general for the small effects in most replicates, data not shown). However the most marked difference between ESSi and SSS is present in Ex5: as for ESSi, SSS misses three effects of "model 1" but in addition $\beta_4 = 1$, $\beta_7 = -1$ and $\beta_8 = 1.5$ receive also very low marginal posterior probability, red right triangle, with high variability across replicates, vertical red dashed line. Moreover on the extreme left, as noted before, ESSi is able to capture the biggest coefficient of "model 2" while SSS misses completely all contaminated effects. No noticeable differences between ESSi and SSS are present in Ex1 and Ex2 for the marginal posterior probability, while in Ex4, SSS shows more variability in $p\left(\gamma_j = 1\,|y\right)$ (red dashed vertical lines compared to blue solid vertical lines) for some covariates that do receive the highest marginal posterior probability.

In contrast to the differences in the marginal posterior probability of inclusion, there is general agreement between the two algorithms with respect to some measures of goodness of fit and stability, see Table 3. Again, not surprisingly, the main difference is seen in Ex5 where ESSi with $\tau = 1$ reaches a better R^2_γ both for the maximum and the $1,000$ largest $p\left(\gamma\,|y\right)$. SSS shows more stability in all examples, but the last: this was somehow expected since one key features of SSS in its ability to move quickly towards the right model and to persist on it (Hans *et al.*, 2007), but a drawback of this is its difficulty to explore far apart models with competing R^2_γ as in Ex5. Note that ESSi shows a small improvement of R^2_γ in all the simulated examples. This is related to the ability of ESSi to pick up some of the small effects that are missed by SSS, see Figure 2. Finally ESSi shows a remarkable superiority in terms of computational time especially when the simulated (and estimated) p_γ is large (in other simulated examples, data not shown, we found this is always true when $p_\gamma \gtrsim 10$): the explanation lies in the number of different models SSS and ESSi evaluate at each sweep. Indeed, SSS evaluates $p + p_\gamma\left(p - p_\gamma\right)$, where p_γ is the size of the current model, while ESSi theoretically analyses an equally large number of models, pL, but, when $p > n$, the actual number of models evaluated is drastically reduced thanks to our FSMH sampler. In only one case SSS beats ESSi in term of computational time (Ex5), but in this instance SSS clearly underestimates the simulated model and hence performs less evaluations than would be necessary to explore faithfully the model space. In conclusion, we see that the rich porfolio of moves and the use of parallel chains makes ESS robust for tackling complex covariate space as well as competitive against a state of the art search algorithm.

[Table 3 about here – Figure 2 about here]

4.4 Performance of ESS with hyperprior on τ

In the previous Section we reported the comparison between ESSi with τ fixed at 1 and SSS. However this is just one over many configurations of our algorithm: several others can be thought of using both g-priors or independent priors with or without a hyperprior on τ. In Figure 3 we illustrate the performance of ESSg when the Zellner-Siow prior (8) is adopted and that of ESSi when a diffuse but proper exponential prior is specified for τ. We stress that this analysis is done purely with the aim to show the behaviour of the proposed algorithm and we do not enter here into the debate of which is the optimal prior for the regression coefficients. Figure 3 (a) illustrate these comparisons on example Ex3. Firstly we note that both ESSi specifications recover well the true model, assigning a small posterior probability of inclusion for the first 15 covariates. However while ESSi shows some uncertainty about the set of predictors not associated to y, ESSg has the remarkable ability to ignore them completely. On the other hand, the uncertainty for ESSg is shifted to the next group of variables whose effect is small, $(\beta_{16}, \ldots, \beta_{30}) = (1, \ldots, 1)$, compared to the simulated error variance: the median of the posterior probability of inclusion averaged across replicates is close to 1 for all of them, but the interquatile range shows non negligible uncertainty about the estimates.

Figure 3 (b) presents the trace plot and the posterior kernel density of τ for one replicate of Ex3 when two different configurations of ESS are adopted, ESSg with the Zellner-Siow prior, top panels, and ESSi with a diffuse exponential prior centred in 10, bottom panels. In both cases *equilibrium* on the product space is easily reached and *marginally* this is evident from the trace plots of τ, left panels. Moreover the chains mix well with an acceptance rate extremely close to the target value, 0.448 and 0.445 respectively and they move quickly, after few iterations, to the target distribution.

The right panels show a complementary story. For ESSg, top right panel, $p\left(\tau \left| y, X\right.\right)$, black solid line, leans quite far apart from the prior distribution, red solid line. The posterior mode is $7,223$, a value almost double with respect to the Benchmark prior proposed by Fernández *et al.* (2001) in the g-prior set-up. Finally the bottom right panel presents the posterior kernel density of the variable selection coefficient obtained running ESSi when a diffuse prior for τ is adopted, red solid line: in this case the posterior mass concentrates around 1.144, a value not very far from 1 which is the recommended choice for τ after standardisation of the covarites.

[Figure 3 about here]

5 Illustration of ESS on real data sets

The first real data example is an application of the Gaussian linear regression model to investigate genetic regulation. To discover the genetic causes of variation in the expression of genes, gene expression data are treated as a quantitative phenotype while genotype data (SNPs) are used as predictors. This analysis, known as expression Quantitative Trait Loci (eQTL), can be seen as an extension of genetic mapping, i.e. locating the source of variation of quantitative traits. Two main problems are related to eQTL analysis: the large number of responses which are not independent and the even larger number of predictors. In current practice, the first problem is ignored, while the second is traditionally solved using univariate measures of association, e.g. a t-test, corrected for multiplicity. Such analyses ignore the interesting possibility of control by more than one gene, so called polygenic control.

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 data since environmental sources of variation are taken under control as well as the biological diversity of the sample. For illustration, we report the analysis of one gene expression response, where we both apply ESSi with and without the hyperprior on τ (see Table 4, eQTL analysis). In this latter case, thanks to the adaptive proposal, the Markov Chain for τ mixes very well reaching an overall acceptance rate which is close to the target value 0.44. Moreover, despite the flat prior that we placed on τ, the posterior distribution is concentrated around 1.430.

In both cases a good mixing among the $L = 4$ chains is obtained. The automatic tuning of the temperature ladder works quite well in both cases reaching an acceptance rate for the monitored exchange operator very close to the optimal one, 0.50. Other measures of mixing show no erratic behaviour of the chains with the last one, $l = 4$ interacting marginally less with the others. Finally running ESS on the same 2MHz CPU with 1.5 Gb RAM desktop computer for $20,000$ sweeps after a burn-in of $5,000$, the computational time is rather similar with or without the hyperprior τ, 28 and 30 minutes respectively.

The main difference among the two implementations of ESSi is related to the posterior model size. When τ is fixed, there is more uncertainty and consequently, more support for larger models, see Figure 4 (a). This is reflected in the marginal posterior probability of inclusion: while there is general agreement for the larger effects, ESSi with a diffuse prior shrink more the small effects. This is not surprising considering the value of the posterior mean of the variable selection coefficient 1.430 greater than 1. On the other hand the best model visited is the same for both version of ESSi, while, when a hyperprior on τ is implemented, we observe a clear degradation of the stability index. This is also evident by looking at "$\overline{R_\gamma^2}$: $1,000$ largest $p\left(\gamma\,|y\,\right)$", see Table 4 eQTL analysis. In both cases we fix $E\left(p_\gamma\right) = 5$ and $V\left(p_\gamma\right) = 3$.

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 Spectography 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 normalised the signal. Since the correlation structure of the SNPs reflects rich patterns of linkage disequilibrium, we decided to apply ESSg, the version of our algorithm with g-priors that is better suited to complex correlation structures among the predictors, coupled with the Zellner-Siow prior (8).

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 4, mQTL analysis). Note that the posterior mean of τ, 89.810, is not close to commonly chosen values for τ, for example the Unit Information Prior $\tau = n$ or the Benchmark prior $\tau = p^2$ (Fernández *et al.*, 2001). In this complex example, the data driven level of shrinkage plays a fundamental role to select the parsimonious set of predictors. In both examples, the posterior model size support polygenic control (Figure 4) highlighting the interest of performing multivariate analysis in genomics.

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 4 by the stability index (≈ 0), we believe that the number of iterations chosen really exceed what it 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 4 about here – Figure 4 about here]

6 Discussion

6.1 MCMC issues

The key idea in constructing an effective MCMC sampler for γ and τ is to add an extra parameter, the temperature, that weakens the likelihood contribution and allows to escape 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 to be exchanged, 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. We believe that using parallel tempering in this way has the potential to be effective in a wide range of situations where the posterior space is multimodal. In order for our algorithm to be able to tackle the case where p is large with respect to p_γ, the second important element is the use of a Metropolised Gibbs sampling-like step restricted to a well chosen set of indices in the local updating of the latent binary vector, rather than an MC^3 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 when p is large and vanilla add/delete moves as outlined in Subsection 4.2.

Note that whilst we found necessary to develop an extensive and diverse set of moves to update γ, if a model with a prior on τ is preferred, the updating of τ 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 Subsection 4.4 and Section 5. Finally, note that 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 by using suitably reweighting of the output of all the chains, in the spirit of Gramacy *et al.* (2007) is an area for further exploration, which is beyond the scope of the present work.

6.2 Extensions

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, 2002). Much of our set-up and algorithm carries through without difficulties and we implemented our algorithm in this framework in a challenging case study in genomics with multidimensional outcomes (four phenotypes and $1,421$ genotypes).

Another possible extension is to analyse multidimensional response Y by a system of q regressions linked in a suitable hierarchical way once α, β_γ and σ^2 have been integrated out. This is a promising area of research that we have already started and are keen to fully pursue in near future.

Appendix

A FSMH schemes

Let $\gamma_{l,j}$, $l = 1, \ldots, L$ and $j = 1, \ldots, p$ to denote the jth latent binary indicator in the lth chain. As in Kohn *et al.* (2001), let $\gamma_{l,j}^{(1)} = (\gamma_{l,1}, \ldots, \gamma_{l,j-1}, \gamma_{l,j} = 1, \gamma_{l,j+1}, \ldots, \gamma_{l,p})^T$ and $\gamma_{l,j}^{(0)} = (\gamma_{l,1}, \ldots, \gamma_{l,j-1}, \gamma_{l,j} = 0, \gamma_{l,j+1}, \ldots, \gamma_{l,p})^T$. Furthermore let $L_{l,j}^{(1)} \propto p\left(y \left| \gamma_{l,j}^{(1)}, g\right.\right)$ and $L_{l,j}^{(0)} \propto p\left(y \left| \gamma_{l,j}^{(0)}, g\right.\right)$ and finally $\theta_{l,j}^{(1)} = p\left(\gamma_{l,j} = 1 \left| \gamma_{l,j^-}\right.\right)$ and $\theta_{l,j}^{(0)} = 1 - \theta_{l,j}^{(1)}$. From (6) it is easy to prove that

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

where p_{γ_l} is the current model size for the lth individual. Using the above equation, for $\gamma_{l,j} = 1$ the normalised version of (15) can be written as

$$p\left(\gamma_{l,j} = 1 \left| y, \gamma_{l,j^-}, g\right.\right)^{1/t_l} = \frac{\theta_{l,j}^{(1)\ 1/t_l} L_{l,j}^{(1)\ 1/t_l}}{S\left(1/t_l\right)}, \tag{A.2}$$

where $S\left(1/t_l\right) = \theta_{l,j}^{(1)\ 1/t_l} L_{l,j}^{(1)\ 1/t_l} + \theta_{l,j}^{(0)\ 1/t_l} L_{l,j}^{(0)\ 1/t_l}$ with $p\left(\gamma_{l,j} = 0 \left| y, \gamma_{l,j^-}, g\right.\right)^{1/t_l}$ defined similarly. Hence if $\theta_{l,j}^{(1)\ 1/t_l}$ is very small, then $p\left(\gamma_{l,j} = 1 \left| y, \gamma_{l,j^-}, g\right.\right)^{1/t_l}$ is small as well.

In the following we derive two FSMH schemes specialised for EMC/PT. Omitting the subscript l for simplicity, we define $Q\left(1 \to 0\right) = Q\left(\gamma_{l,j}^{(1)} \to \gamma_{l,j}^{(0)}\right)$ as the proposal probability to go from 0 to 1 and $Q\left(1 \to 0\right)$ the proposal probability to go from 1 to 0. Moreover using the notation introduced before, the Metropolised version of (15) to go from 0 to 1 in the EMC local move is

$$\alpha_l^{\mathrm{MH}}\left(0 \to 1\right) = \min\left\{1, \frac{\theta_{l,j}^{(1)\ 1/t_l} L_{l,j}^{(1)\ 1/t_l}}{\theta_{l,j}^{(0)\ 1/t_l} L_{l,j}^{(0)\ 1/t_l}} \frac{Q\left(1 \to 0\right)}{Q\left(0 \to 1\right)}\right\} \tag{A.3}$$

with a similar expression for $\alpha_l^{\mathrm{MH}}\left(1 \to 0\right)$. 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) ${L_{l,j}^{(0)}}^{1/t_l} \Big/ {L_{l,j}^{(1)}}^{1/t_l} \geq 1\,;$

b) ${L_{l,j}^{(1)}}^{1/t_l} \Big/ \tilde{S}\left(1/t_l\right) \geq 1\,;$

c) ${L_{l,j}^{(0)}}^{1/t_l} \Big/ \tilde{S}\left(1/t_l\right) < 1\,,$

where $\tilde{S}\left(1/t_l\right) = S\left(1/t_l\right) \Big/ \left({\theta_{l,j}^{(1)}}^{1/t_l} + {\theta_{l,j}^{(0)}}^{1/t_l}\right).$

Sampling scheme 1

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

- $Q^{FSMH_1}\left(0 \to 1\right) = \tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right) \min\left\{1,\, {L_{l,j}^{(1)}}^{1/t_l} \Big/ \tilde{S}\left(1/t_l\right)\right\}$

- $Q^{FSMH_1}\left(1 \to 0\right) = \tilde{\theta}_{l,j}^{(0)}\left(1/t_l\right) \min\left\{1,\, {L_{l,j}^{(0)}}^{1/t_l} \Big/ \tilde{S}\left(1/t_l\right)\right\}$

with $\tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right) = {\theta_{l,j}^{(1)}}^{1/t_l} \Big/ \left({\theta_{l,j}^{(1)}}^{1/t_l} + {\theta_{l,j}^{(0)}}^{1/t_l}\right)$ *and* $\tilde{\theta}_{l,j}^{(0)}\left(1/t_l\right) = 1 - \tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right)$*. Then the acceptance probabilities are*

$$\alpha_l^{FSMH_1}\left(0 \to 1\right) = \begin{cases} 1 & \text{if } {L_{l,j}^{(1)}}^{1/t_l} \Big/ {L_{l,j}^{(0)}}^{1/t_l} \geq 1 \\ \tilde{S}\left(1/t_l\right) \Big/ {L_{l,j}^{(0)}}^{1/t_l} & \text{if } {L_{l,j}^{(1)}}^{1/t_l} \Big/ {L_{l,j}^{(0)}}^{1/t_l} < 1 \end{cases}$$

$$\alpha_l^{FSMH_1}\left(1 \to 0\right) = \begin{cases} 1 & \text{if } {L_{l,j}^{(0)}}^{1/t_l} \Big/ {L_{l,j}^{(1)}}^{1/t_l} \geq 1 \\ \tilde{S}\left(1/t_l\right) \Big/ {L_{l,j}^{(1)}}^{1/t_l} & \text{if } {L_{l,j}^{(0)}}^{1/t_l} \Big/ {L_{l,j}^{(1)}}^{1/t_l} < 1 \end{cases}$$

The above sampling scheme is implemented as follows. For a given l and for $j = 1, \ldots, p$ (or any permutation of them) let $u \sim U\left(0, 1\right)$. Consider for simplicity $0 \to 1$. If $u > \tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right)$ then $u > Q^{\mathrm{FSMH}_1}\left(0 \to 1\right)$ and the move is rejected. If $u \leq \tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right)$ then $Q^{\mathrm{FSMH}_1}\left(0 \to 1\right)$ must be evaluated. In the second step the proposal distribution equals (A.2): therefore sampling scheme 1 can be seen as a random scan Metropolised Gibbs sampling where the number of evaluations is linked to the prior/current model size and the temperature attached to the chain.

Sampling scheme 2

The second sampling scheme retains the idea of a two-step Metropolis-Hastings acceptance rate as in FSMH$_1$. However it simplifies ever further the computation requirements using the normalised tempered version of (6) as a proposal.

Proposition 3 *Let $l = 1, \ldots, L$, $j = 1, \ldots, p$ (or any permutation of them) and*

- $Q^{FSMH_2}\left(0 \to 1\right) = \tilde{\theta}_{l,j}^{(1)}\left(1/t_l\right)$

- $Q^{FSMH_2}\left(1 \to 0\right) = \tilde{\theta}_{l,j}^{(0)}\left(1/t_l\right)$

with $\tilde{\theta}^{(0)}_{l,j}\left(1/t_l\right) = 1 - \tilde{\theta}^{(1)}_{l,j}\left(1/t_l\right)$*. The acceptance probabilities are*

$$\alpha_l^{FSMH_2}\left(0 \to 1\right) = \begin{cases} 1 & \text{if } {L^{(1)}_{l,j}}^{1/t_l} \Big/ {L^{(0)}_{l,j}}^{1/t_l} \geq 1 \\ {L^{(1)}_{l,j}}^{1/t_l} \Big/ {L^{(0)}_{l,j}}^{1/t_l} & \text{if } {L^{(1)}_{l,j}}^{1/t_l} \Big/ {L^{(0)}_{l,j}}^{1/t_l} < 1 \end{cases}$$

$$\alpha_l^{FSMH_2}\left(1 \to 0\right) = \begin{cases} 1 & \text{if } {L^{(0)}_{l,j}}^{1/t_l} \Big/ {L^{(1)}_{l,j}}^{1/t_l} \geq 1 \\ {L^{(0)}_{l,j}}^{1/t_l} \Big/ {L^{(1)}_{l,j}}^{1/t_l} & \text{if } {L^{(0)}_{l,j}}^{1/t_l} \Big/ {L^{(1)}_{l,j}}^{1/t_l} < 1 \end{cases}$$

The following Proposition compares the efficiency of the tempered Gibbs sampling (15) and the proposed FSMH schemes in Proposition 2 and 3. Proof is easy to check and it is omitted.

Proposition 4 *Let* $Q^G\left(\cdot\right)$, $Q^{FSMH_1}\left(\cdot\right)$ *and* $Q^{FSMH_2}\left(\cdot\right)$ *as proposal probabilities and* $\alpha^{FSMH_1}\left(\cdot\right)$ *and* $\alpha^{FSMH_2}\left(\cdot\right)$ *as acceptance probabilities for Gibbs sampling and FSMH schemes respectively, then*

$$\begin{aligned} Q^G\left(0 \to 1\right) &> Q^{FSMH_1}\left(0 \to 1\right)\alpha^{FSMH_1}\left(0 \to 1\right) \\ &= Q^{FSMH_2}\left(0 \to 1\right)\alpha^{FSMH_2}\left(0 \to 1\right) \end{aligned}$$

$$\begin{aligned} Q^G\left(1 \to 0\right) &> Q^{FSMH_1}\left(1 \to 0\right)\alpha^{FSMH_1}\left(1 \to 0\right) \\ &= Q^{FSMH_2}\left(1 \to 0\right)\alpha^{FSMH_2}\left(1 \to 0\right) \end{aligned}$$

The above Proposition states 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

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.D. and Donnelly, P. (2005). A haplotype map of the human genome. *Nature*, **437**, 1299-1320.

Bae, N. and Mallick, B.K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423-3430.

Berger, J.O. and Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statist. Neerl.*, **59**, 3-15.

Berger, J.O., Liseo, B. and Wolpert, R.L. (1999). Integrating likelihood methods for eliminating nuisance parameters. *Statist. Sci.*, **14**, 1-28.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: Wiley.

Brown, P.J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc. B*, **60**, 627-641.

Brown, P.J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *J. R. Statist. Soc. B*, **64**, 519-536.

Calvo, F. (2005) All-exchange parallel tempering. *J. Chem. Phys.*, **123**, 1-7.

Celeux, G., Marin, J.-M. and Robert, C.P. (2006). Sélection Bayésienne de variables en régression lineaire. *J. Soc. Franc. Statist.*, **147**, 59-79.

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.

Cripps, E., Kohn, R. and Nott, D. (2006). Bayesian subset selection and model averaging using a centred and dispersed prior for the error variance. *Aust. N. Z. J. Stat.*, **48**, 237-252.

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.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998). Automatic Bayesian curve fitting. *J. R. Statist. Soc. B*, **60**, 333-350.

Denison, D.G.T., Holmes, C.C., Mallick, B.K. and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Wiley.

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 Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731-747.

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. Claredon 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. Available at the WEB site: `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., Wallace, C.A., Zimdahl, H., Petretto, E. and Aitman, T.J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.*, **37**, 243-253.

Hukushima, K. and Nemoto, K. (1996). Exchange Monte Carlo methods and application to spin glass simulations. *J. Phys. Soc. Jpn.*, **65**, 1604-1608.

Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statist. Comp.*, **11**, 313-322.

Iba, Y. (2001). Extended Ensemble Monte Carlo. *Int. J. Mod. Phys., C*, **12**, 623-56.

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.

Liang, F. and Wong, W.H. (2001). Real-parameter evolutionary Monte Carlo with application to Bayesian mixture models. *J. Am. Statist. Assoc.*, **96**, 653-666.

Liu, J.S. (1996). Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, **83**, 681-682.

Liu, J.S. (2001). *Monte Carlo strategies in scientific computations.* Springer: New York.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Int. Statist. Rev.*, **63**, 215-232.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europh. Lett.*, **19**, 451-458.

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 Swedsen-Wang algorithm. *J. Comp. Graph. Statist.*, **13**, 141-157.

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

Roberts, G.O. and Rosenthal, J.S. (2007). Example of adaptive MCMC. Tech. rep., Dept. of Statistics, University of Toronto. Available at the WEB site: `http://www.probability.ca/jeff/research.html`

Scott, J.G. and Berger, J.O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Stat. Plan. Inf.*, **136**, 2144-2162.

Sha, N., Vannucci, M., Tadesse, M., Brown, P., Dragoni, I., Davies, N., Roberts, T., Contestabile, A., Salmon, N., Buckley, C. and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**, 812-819.

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.

	Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
$E\left(p_\gamma\right)$	5	10	20	5	5	5	5	5
Add/delete	0.036	0.054	0.098	0.066	0.086	-	-	-
	(0.016)	(0.017)	(0.023)	(0.020)	(0.031)	-	-	-
Swap	0.063	0.100	0.165	0.070	0.106	-	-	-
	(0.015)	(0.019)	(0.022)	(0.015)	(0.053)	-	-	-
Crossover	0.249	0.270	0.271	0.157	0.215	0.147	0.170	0.193
	(0.021)	(0.029)	(0.036)	(0.018)	(0.022)	(0.028)	(0.023)	(0.028)
Exchange	0.500	0.493	0.500	0.582	0.492	0.517	0.505	0.497
(delayed rejection)	(0.040)	(0.043)	(0.040)	(0.020)	(0.071)	(0.105)	(0.013)	(0.072)

Table 1: Mean and standard deviation in brackets of EMC acceptance rates across replicates for ESS*i* with $\tau = 1$.

		Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
$E\left(p_\gamma\right)$		5	10	20	5	5	5	5	5
Swapping	$l = 1$	0.157	0.137	0.110	0.065	0.160	0.180	0.201	0.214
	$l = 2$	0.250	0.232	0.204	0.185	0.271	0.276	0.300	0.316
	$l = 3$	0.220	0.220	0.223	0.255	0.245	0.223	0.231	0.231
	$l = 4$	0.240	0.252	0.280	0.293	0.215	0.206	0.182	0.167
	$l = 5$	0.142	0.160	0.182	0.201	0.110	0.112	0.083	0.070
Overlapping	$l = 1, 2$	1.360	1.600	2.101	2.680	1.350	0.733	0.569	0.526
	$l = 2, 3$	1.570	1.570	1.600	0.870	1.430	1.021	0.913	0.893
	$l = 3, 4$	1.400	1.290	1.050	0.600	2.111	1.329	1.491	1.696
	$l = 4, 5$	1.100	0.992	0.690	1.251	4.131	1.503	2.304	2.499

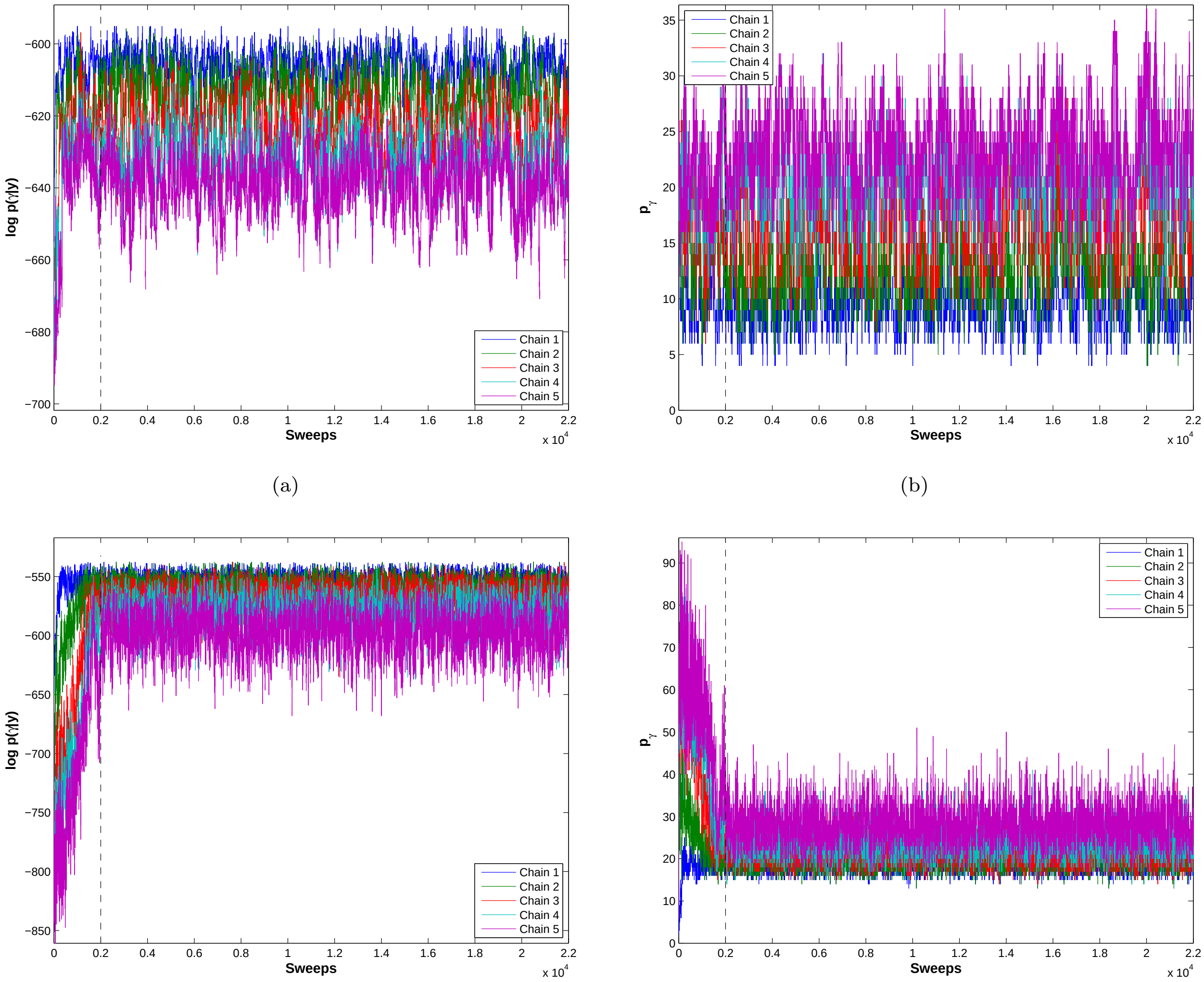
Table 2: Swapping probability for ESS*i* with $\tau = 1$ defined as the observed frequency of successful swaps for each chain (including delayed rejection exchange and all-exchange operators) averaged across replicates. Overlapping measure defined as $V\left(f\left(\gamma_l\right)\right)\left(1/t_{l+1} - 1/t_l\right)^2$, Liang and Wong (2000) with $f\left(\gamma_l\right) = \log p\left(y\left|\gamma_l\right.\right) + \log p\left(\gamma_l\right)$. Target value for consecutive chains is $O\left(1\right)$.

		Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
	$E\left(p_{\gamma}\right)$	5	10	20	5	5	5	5	5
ESSi, τ=1	R_{γ}^{2}: $\max p\left(\gamma\,|y\right)$	0.864	0.867	0.871	0.975	≈ 1	0.962	0.703	0.997
		(0.029)	(0.027)	(0.023)	(0.003)	(≈ 0)	(0.011)	(0.043)	(0.005)
	$\overline{R_{\gamma}^{2}}$: $1,000$ largest $p\left(\gamma\,|y\right)$	0.863	0.866	0.874	0.975	≈ 1	0.957	0.689	0.997
		(0.027)	(0.026)	(0.023)	(0.003)	(≈ 0)	(0.014)	(0.048)	(0.003)
	Stability	0.003	0.003	0.005	≈ 0	(≈ 0)	0.005	0.015	0.002
		(0.001)	(0.002)	(0.002)	(≈ 0)	(≈ 0)	(0.004)	(0.007)	(0.002)
	Time	6	6	7	16	18	166	338	202
		(< 1)	(< 1)	(< 1)	(< 1)	(1)	(32)	(43)	(40)
SSS	R_{γ}^{2}: $\max p\left(\gamma\,|y\right)$	0.863	0.867	0.870	0.975	≈ 1	0.956	0.577	0.997
		(0.027)	(0.025)	(0.024)	(0.003)	(≈ 0)	(0.016)	(0.074)	(0.004)
	$\overline{R_{\gamma}^{2}}$: $1,000$ largest $p\left(\gamma\,|y\right)$	0.863	0.867	0.870	0.975	0.999	0.955	0.565	0.996
		(0.027)	(0.025)	(0.024)	(0.003)	(≈ 0)	(0.016)	(0.078)	(0.004)
	Stability	0	0	≈ 0	≈ 0	≈ 0	0.001	0.009	0.004
		(0)	(0)	(≈ 0)	(≈ 0)	(≈ 0)	(0.002)	(0.015)	(0.006)
	Time	12	12	13	118	497	502	169	549
		(1)	(2)	(2)	(26)	(75)	(241)	(81)	(159)

Table 3: Comparison between ESSi with $\tau = 1$ and SSS for the six simulated examples. Standard deviation in brackets.

		Mode$(p_{\gamma}\,|y\,)$	$E\left(\tau\,|y\,\right)$	R_{γ}^{2}: $\max p\left(\gamma\,|y\,\right)$	$\overline{R_{\gamma}^{2}}$: $1,000$ largest $p\left(\gamma\,|y\,\right)$	Stability
eQTL	ESSi, $\tau = 1$	3	–	0.859	0.850	0.054
	ESSi with $p\left(\tau\right)^{*}$	3	1.430	0.859	0.764	0.110
mQTL	ESSg with $p\left(\tau\right)^{**}$	2	89.810	0.843	0.843	≈ 0
		Crossover	Exchange (delayed rejction)	Acceptance rate τ	Time	
eQTL	ESSi, $\tau = 1$	0.078	0.548	–	27	
	ESSi with $p\left(\tau\right)^{*}$	0.123	0.525	0.452	30	
mQTL	ESSg with $p\left(\tau\right)^{**}$	0.080	0.628	0.452	309	

Table 4: Comparison between ESSi with and without the prior on τ for the first real data example, eQTL analysis, and ESSg with the Zellner-Siow prior for the second example, mQTL analysis ($p\left(\tau\right)^{*} = \mathcal{E}xp\left(10\right)$ and $p\left(\tau\right)^{**} = InvGam\left(1/2, n/2\right)$).



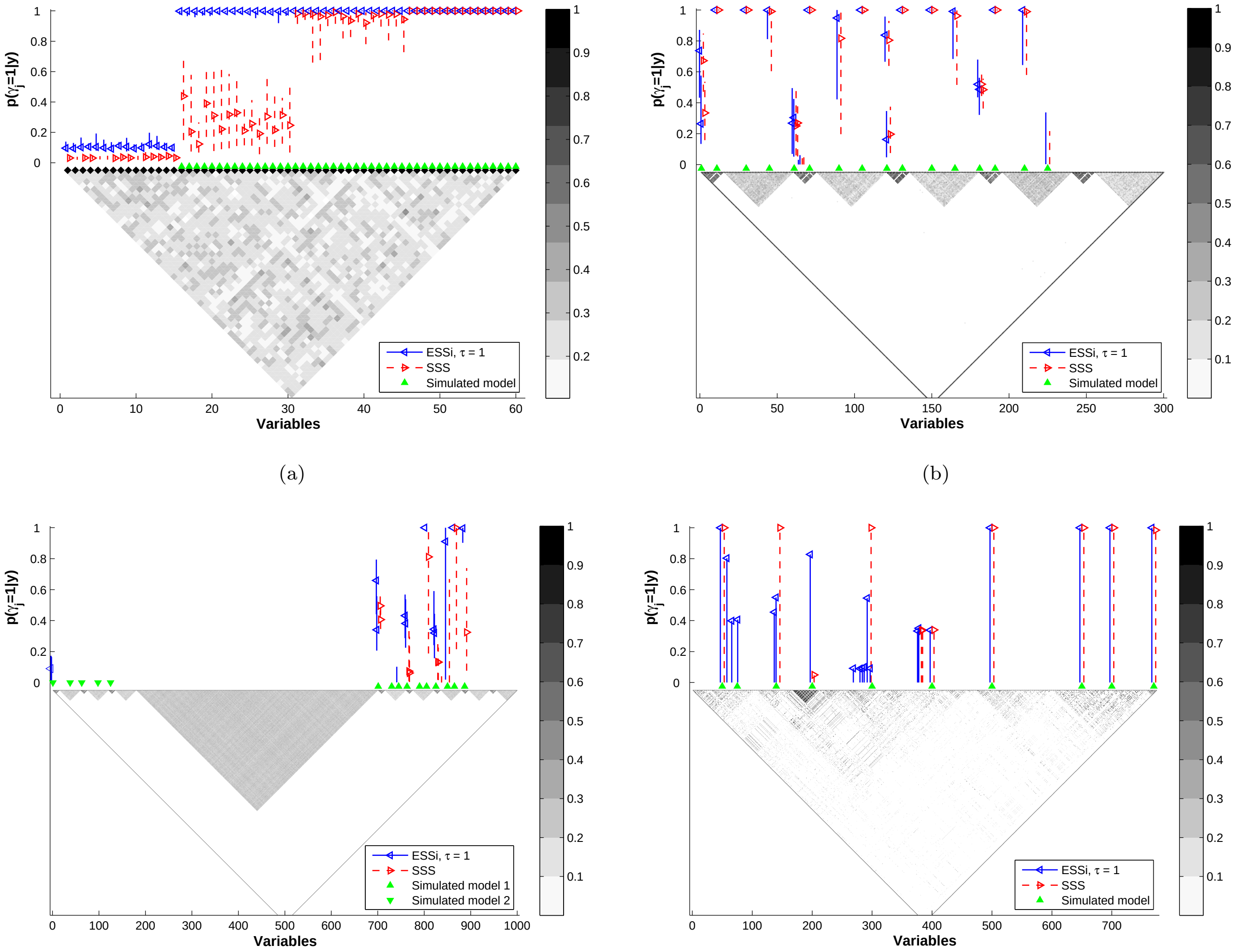
(a)

(b)

(c)

(d)

Figure 1: For ESSi with $\tau = 1$: (a) trace plot of the log posterior probability, $\log p\left(\gamma\,|y\right)$ and (b) model size, p_γ, across sweeps for one replicate of Ex1 with $E\left(p_\gamma\right) = 20$, top panels and Ex4, bottom panels. Vertical dashed lines indicate the end of the burn-in.



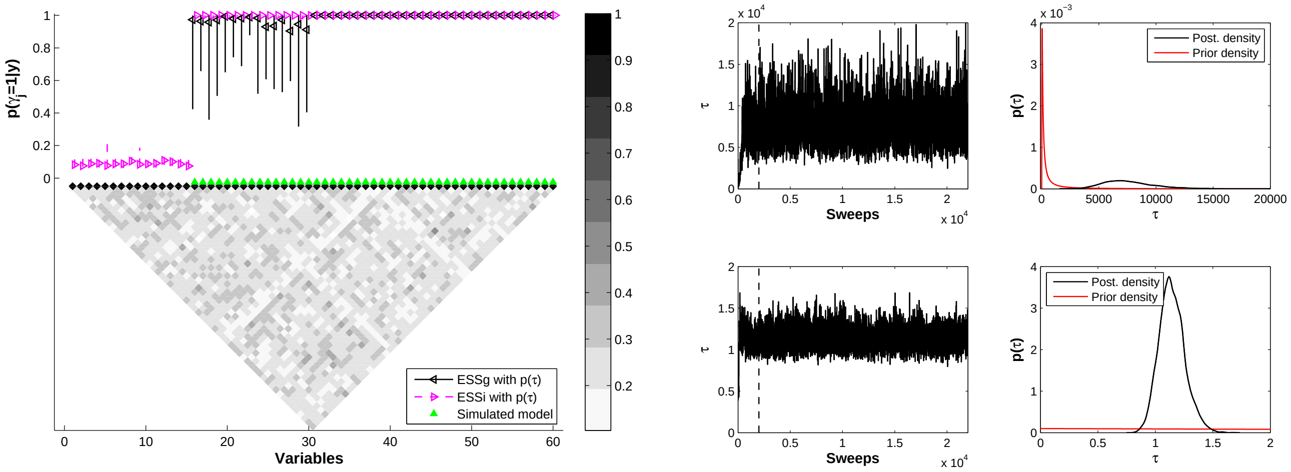
(a)

(b)

(c)

(d)

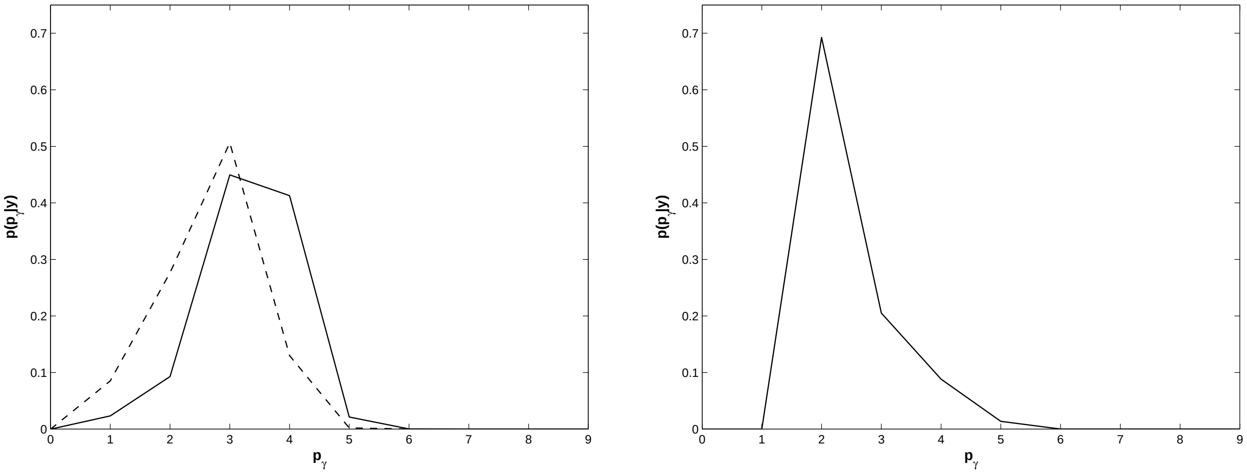
Figure 2: Median and interquantile range of the marginal posterior probability of inclusion (19) for Ex3, (a), Ex4, (b) and Ex5, (c), across replicates. Each graph is constructed as follows: bottom part, pairwise r^2 for one selected replicate, grey scale indicates different values of squared correlation; blue left and red right triangles, median of $p\left(\gamma_j = 1\,|y\right)$, $j = 1, \ldots, p$, across replicates for ESSi with $\tau = 1$ and SSS respectively; vertical blue solid lines and vertical red dashed lines, interquantile range of $p\left(\gamma_j = 1\,|y\right)$, $j = 1, \ldots, p$, across replicates for ESSi and SSS respectively; upper and lower green triangles, simulated models. Selected replicate of Ex6, (d), shows marginal posterior probability of inclusion (blue left and red right triangles for ESSi $\tau = 1$ and SSS respectively). Marginal posterior probability of inclusion lower than 0.025 not shown.



(a)

(b)

Figure 3: (a) Median and interquantile range of the marginal posterior probability of inclusion (19) across replicates for Ex3 when ESSg is applied with Zellner-Siow prior (median, back left triangles and interquantile range, vertical black solid lines) and ESSi is used with a proper but diffuse exponential prior centred in 10 (median, magenta right triangles and interquantile range, vertical magenta dashed lines). Upper green triangles, simulated models. Marginal posterior probability of inclusion lower than 0.025 not shown. (b) Top panels, trace plot and posterior kernel density of τ for ESSg with Zellner-Siow prior; bottom panels, trace plot and posterior kernel density of τ for ESSi with diffuse exponential prior centred in 10. Vertical dashed lines on the right panels indicate the end of the burn-in. Red lines on the right panels show prior density.



(a)

(b)

Figure 4: (a) Posterior model size for the first real data example related to eQTL analysis: black solid line for ESSi with τ fixed at 1 and black dashed line for ESSi with flat exponential distribution on τ. (b) Posterior model size for mQTL analysis, second real data example, using ESSg coupled with Zellner-Siow prior.