Supplementary material:

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

A Technical details of EMC implementation

In this Section we will describe some technical details omitted from the paper and related to the sampling schemes we used for the population of binary latent vectors $\boldsymbol{\gamma}$ and the selection coefficient τ.

A.1 EMC sampler for $\boldsymbol{\gamma}$ (see Section 3.1 in the paper)

Global move: exchange operator

The exchange operator can be seen as an extreme case of crossover operator, where the first proposed chain receives the whole second chain state $\gamma_l' = \gamma_r$, and the second proposed chain receives the whole first state chain $\gamma_r' = \gamma_l$, respectively.

In order to achieve a good acceptance rate, the exchange operator is usually applied on adjacent chains in the temperature ladder, which limits its capacity for mixing. To obtain better mixing, we implemented two different approaches: the first one is based on Jasra *et al.* (2007) and the related idea of delayed rejection (Green and Mira, 2001); the second one on Gibbs distribution over all possible chains pairs (Calvo, 2005).

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). Details are reported below.

Suppose two chains are selected at random, l and r with $l \neq r$, in order to swap their binary latent vector. Then, given that $\gamma_l' = \gamma_r$, $\gamma_r' = \gamma_l$ and $Q_t(\boldsymbol{\gamma} \to \boldsymbol{\gamma}') = Q_t(\boldsymbol{\gamma}' \to \boldsymbol{\gamma})$, (13) reduces to

$$\alpha_1(\boldsymbol{\gamma} \to \boldsymbol{\gamma}') = \min\left\{1, \frac{\exp\{f(\gamma_r|\tau)/t_l + f(\gamma_l|\tau)/t_r\}}{\exp\{f(\gamma_l|\tau)/t_l + f(\gamma_r|\tau)/t_r\}}\right\}.$$

Since the two chains are selected at random, the above acceptance probability decreases exponentially with the difference $|1/t_l - 1/t_r|$ and therefore most of the proposed moves are rejected. If rejected, a delayed rejection-type move is applied between two random adjacent chains, with l the first one and s, $|l - s| = 1$, the second one, giving rise to the new acceptance probability

$$\alpha_2(\boldsymbol{\gamma} \to \boldsymbol{\gamma}'') = \min\left\{1, \frac{\exp\{f(\gamma_s|\tau)/t_l + f(\gamma_l|\tau)/t_s\}}{\exp\{f(\gamma_l|\tau)/t_l + f(\gamma_s|\tau)/t_s\}} \frac{1 - \alpha_1(\boldsymbol{\gamma}'' \to \boldsymbol{\gamma}^*)}{1 - \alpha_1(\boldsymbol{\gamma} \to \boldsymbol{\gamma}')}\right\},$$

where the pseudo move $\boldsymbol{\gamma}^*$ is necessary in order to maintain the detailed balance condition (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 $L(L-1)/2$ exchange acceptance rates between all chains pairs, inclusive the rare ones, $\tilde{p}_{l,r} = \exp\{(f(\gamma_r|\tau) - f(\gamma_l|\tau))(1/t_l - 1/t_r)\}$. 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

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

where in (S.1) 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

First we select the number L of chains close to the complexity of the problem, i.e. $E(p_\gamma)$, although the size of the data and computational limits need to be taken into account. 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 (2008), 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 close 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 have 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 total number of batches in the burn-in period.

A.2 Adaptive Metropolis-within-Gibbs for τ (see Section 3.2 in the paper)

Laplace approximation for the conditional marginal likelihood

Under model (1) and prior specification for α, (2) and (3), we provide the Laplace approximation of $p(y|\gamma, \tau)$ for the g-prior case, while the approximation for the independent case can be derived following the same line of reasoning. For easy of notation we drop the chain subscript index and we assume that the observed responses y have been centred with mean 0, i.e. $(y - \bar{y}_n) \equiv y$. In the following we will distinguish the cases in which the posterior mode $\hat{\tau}_\gamma$ is a solution of a cubic or quadratic equation. Conditions on the existence of the solutions are provided as well as those that guarantee the positive semidefiniteness of the variance approximation. Recall that

$$p(y|\gamma) = \int \exp\{\log(p(y|\gamma, \tau)\, p(\tau))\}\, d\tau$$

$$\approx \quad \sqrt{2\pi}\sigma_{\hat{\lambda}}\left(\log p\left(y\left|\gamma,\hat{\lambda}\right.\right)+\log p\left(\hat{\lambda}\right)+\log J\left(\hat{\lambda}\right)\right),$$

where $\hat{\lambda}$ is the posterior mode after the transformation $\lambda=\log(\tau)$, which is necessary to avoid problems on the boundary, $\sigma_{\hat{\lambda}}$ is the approximate squared root of the variance calculated in $\hat{\lambda}$ and $J(\cdot)$ is the Jacobian of the transformation. Details about Laplace approximation can be found in Tierney and Kadane (1986). Similar derivations when $p\left(\sigma^2\right)\propto\sigma^{-2}$ are presented in Liang *et al.* (2008). Finally throughout the presentation we will assume that $n>p_\gamma$ and that a_g and b_g are fixed small as in Kohn *et al.* (2001).

Cubic equation for Zellner-Siow priors

If $p(\tau)=InvGa(a_\tau,b_\tau)$ the posterior $\hat{\lambda}$ mode is the only positive root of the integrand function

$$I_\lambda=\left(1+e^\lambda\right)^{(2a_\sigma+n-1-p_\gamma)/2}\left\{2b_\sigma\left(1+e^\lambda\right)+y^Ty\left[1+e^\lambda\left(1-R_\gamma^2\right)\right]\right\}^{-(2a_\sigma+n-1)/2}\frac{e^{-b_\tau/e^\lambda}}{\left(e^\lambda\right)^{a_\tau+1}}e^\lambda,$$

where the last factor in the above equation $e^\lambda=\left|de^\lambda/d\lambda\right|$ is the Jacobian of the transformation. After the calculus of the first derivative of the log transformation and some algebra manipulations, it can be shown that $e^{\hat{\lambda}}$ is the solution of the cubic equation

$$e^{3\lambda}+\frac{c_1c_3-c_2c_4-(c_3+c_4)a_\tau+c_4b_\tau}{(c_1-c_2-a_\tau)c_4}e^{2\lambda}+\frac{-c_3a_\tau+(c_3+c_4)b_\tau}{(c_1-c_2-a_\tau)c_4}e^\lambda+\frac{c_3b_\tau}{(c_1-c_2-a_\tau)c_4}=0 \quad \text{(S.2)}$$

and that

$$\begin{aligned}\sigma_{\hat{\lambda}}^2 &= -\frac{1}{\left(\log p(y|\gamma,\lambda)+\log p(\lambda)\right)''}\Bigg|_{\lambda=\hat{\lambda}}\\ &= \left[-c_1\frac{e^\lambda}{\left(1+e^\lambda\right)^2}+c_2\frac{c_3c_4e^\lambda}{\left(c_3+c_4e^\lambda\right)^2}+\frac{b_\tau}{e^\lambda}\right]_{\lambda=\hat{\lambda}}^{-1},\end{aligned} \quad \text{(S.3)}$$

where $c_1=(2a_\sigma+n-1-p_\gamma)/2$, $c_2=(2a_\sigma+n-1)/2$, $c_3=2b_\sigma+y^Ty$ and $c_4=2b_\sigma+y^Ty\left(1-R_\gamma^2\right)$. Following Liang *et al.* (2008), since $\lim_{\lambda\to-\infty}\partial I_\lambda/\partial\lambda>0$, because $c_3b_\tau>0$, and $\lim_{\lambda\to\infty}\partial I_\lambda/\partial\lambda<0$, because $(c_1-c_2-a_\tau)c_4<0$, at least one real positive solution exists. Moreover since $-(c_3b_\tau)/(c_1-c_2-a_\tau)c_4>0$, the remaining two real solutions should have the same sign (Abramowitz and Stegun, 1970). A necessary condition for the existence of just one real positive solution is that the summation of all the pairs-products of the coefficients is negative

$$\frac{-c_3a_\tau+(c_3+c_4)b_\tau}{(c_1-c_2-a_\tau)c_4}<0$$

and this happens if $b_\tau/a_\tau > c_3/(c_3+c_4)$. When $R^2_\gamma \to 0$ and thus $c_3 = c_4$, the above condition corresponds to $b_\tau > a_\tau/2$ and when $R^2_\gamma \to 1$, as $c_3/(c_3+c_4) \approx 1$ especially when $y^T y$ is large, which might be expected when n becomes large, the condition is equivalent to $b_\tau > a_\tau$. Therefore it turns out that a sufficient condition for the existence of just one real positive solution in (S.2) is $b_\tau > a_\tau$.

The positive semidefiniteness of the approximate variance can be proved as follows. First of all it is worth noticing that all the terms in (S.3) are of the same order $O_p\left(e^{-\lambda}\right)$. Then, when $R^2_\gamma \to 0$, the positive semidefiniteness is always guaranteed, while when $R^2_\gamma \to 1$, provided that $y^T y$ is large, the middle term in (S.3) tends to zero and the condition is fulfilled if $b_\tau > c_1$.

Quadratic equation for Liang *et al.* (2008) prior

If $p(\tau) \propto (1+\tau)^{-c_\tau}$, with $c_\tau > 0$, $e^{\hat{\lambda}}$ is only the positive root of the integrand function

$$I_\lambda = \left(1+e^\lambda\right)^{(2a_\sigma+n-1-p_\gamma-c_\tau)/2} \left\{2b_\sigma\left(1+e^\lambda\right) + y^T y\left[1+e^\lambda\left(1-R^2_\gamma\right)\right]\right\}^{-(2a_\sigma+n-1)/2} e^\lambda$$

or, after the first derivative of the log transformation, the solution of the quadratic equation

$$\left(c_1^* - c_2 + 1\right) c_4 e^{2\lambda} + \left(c_1^* c_3 - c_2 c_4 + c_3 + c_4\right) e^\lambda + c_3 = 0 \tag{S.4}$$

with $c_1^* = \left[2a_\sigma + n - 1 - (p_\gamma + 2c_\tau)\right]/2$ and c_2, c_3 and c_4 defined as above. The discriminant of the quadratic equation is $\Delta = \left(c_1^* c_3 - c_2 c_4 c_3 + c_3 + c_4\right)^2 - 4\left(c_1^* - c_2 + 1\right) c_4 c_3$ which is always greater than zero and therefore two real roots exist. Since one of them is positive in order to prove that (S.4) admits just one positive solution, it is necessary to show that

$$\frac{-\left(c_1^* c_3 - c_2 c_4 + c_3 + c_4\right) - \Delta^{1/2}}{2\left(c_1^* - c_2 + 1\right) c_4} < 0$$

which is true provided that $\left(c_1^* - c_2 + 1\right) c_4 c_3 < 0$. Moreover the approximate variance can be written as

$$\sigma^2_{\hat{\lambda}} = \left[-c_1^* \frac{e^\lambda}{\left(1+e^\lambda\right)^2} + c_2 \frac{c_3 c_4 e^\lambda}{\left(c_3 + c_4 e^\lambda\right)^2}\right]^{-1}_{\lambda=\hat{\lambda}} \tag{S.5}$$

which is positive semidefinite when $R^2_\gamma \to 0$ if $c_2 > c_1^*$, which is always verified, while, if $R^2_\gamma \to 1$ and $y^T y$ is large, equation (S.5) is not positive unless $p_\gamma + 2c_\tau > 2a_\sigma + n - 1$.

The explicit solution of the posterior mode is also available

$$\hat{\tau}_\gamma = \max\left\{\frac{\left(c_4 - c_3\right)/\left(c_1^* - c_2\right)}{c_4/c_1^*} - 1, 0\right\}$$

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

which corresponds to MLE if $c_\tau = 0$.

Diminishing adaptive and bounded conditions

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 $\log(\tau)$ in the g-prior and independent prior case. The variance of the proposal distribution is controlled as illustrated in Roberts and Rosenthal (2008): every 100 EMC sweeps, the same value of sweeps used in the temperature placement, we monitor the acceptance rate of the Metropolis-within-Gibbs algorithm: if it is lower (higher) than the optimal acceptance rate, i.e. 0.44, a constant $\delta_\tau(k)$ is added (subtracted) 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 $|ls_1 - 5|/\tilde{K}$, where $\tilde{K}$ is the total number of batches in the burn-in period: 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. The *diminishing adaptive condition* is obtained imposing $\delta_\tau(k) = \min\{|ls_1 - 5|/\tilde{K}, k^{-1/2}\}$, where k is the current number of batches, including the burn-in. To ensure the *bounded convergence condition* we follow Roberts and Rosenthal (2008), restricting each ls_k to be inside $[M_1, M_2]$ and we fix them equal to $M_1 = -10$ and $M_2 = 10$ respectively. In practise these bounds do not create any restriction since the sequence of the standard deviations of the proposal distribution stabilises almost immediately, indicating that the transition kernel converges in a bounded number of batches, see Figure S.1.

B Performance of ESS: Real data examples

In this section we include Tables and Figures that are commented in Section 4.1 of the paper.

[Table S.1 about here – Table S.2 about here]

[Figure S.1 about here – Figure S.2 about here]

C Performance of ESS: Simulation study

In this Section we report in details on the performance of ESS in a variety of simulated examples. Main conclusions are summarised in the Section 4.2 of the paper.

Firstly we 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 Section 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 priors (8), Z-S hereafter, and for the latter we will specify a proper but diffuse exponential distribution as suggested by Bae and Mallick (2004).

C.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 two simulated examples in George and McCulloch, G&McC hereafter, (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. In particular the jth column of X_1, indicated as $X_{(1)j}$, is simulated as $X_{(1)j} = X_j^* + Z$, where $X_1^*, \ldots, X_{60}^*$ iid $\sim N_n(0, 1)$ independently form

$Z \sim N_n(0,1)$, inducing a pairwise correlation of 0.5. X_2 is generated as follows: firstly we simulated $Z_1, \ldots, Z_{15}$ iid $\sim N_n(0,1)$ and we set $X_{(2)j} = Z_i + 2Z_j$ for $j = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$ only. To induce strong multicollinearity, we then set $X_{(2)2} = X_{(2)1} + 0.15Z_2$, $X_{(2)4} = X_{(2)3} + 0.15Z_4$, $X_{(2)6} = X_{(2)5} + 0.15Z_6$, $X_{(2)7} = X_{(2)8} + X_{(2)9} - X_{(2)10} + 0.15Z_7$ and $X_{(2)11} = X_{(2)14} + X_{(2)15} - X_{(2)12} - X_{(2)13} + 0.15Z_{11}$. A pairwise correlation of about 0.998 between $X_{(2)j}$ and $X_{(2)j+1}$ for $j = 1, 3, 5$ is introduced and similarly strong linear relationship is present within the sets $\left(X_{(2)7}, X_{(2)8}, X_{(2)9}, X_{(2)10}\right)$ and $\left(X_{(2)11}, X_{(2)12}, X_{(2)13}, X_{(2)14}, X_{(2)15}\right)$.

Then, as in Nott and Green, N&G hereafter, (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 Section 3.3 in the paper, 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 $X_3 = [X_2 X_1^*]$ (dimension 120×60). 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 $\beta_\gamma^1 = (2, -1, 1.5, 1, 0.5, 2, -1, 1.5, 2, -1)$) while y_2 is simulated from "model 2" ($\gamma^2 = (1, 38, 63, 98, 125)$ and $\beta_\gamma^2 = (2, -1, 1.5, 1, 0.5)$). 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 Section 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, Section 3.1 in the paper. 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 $L = 5$ with the starting value of $\boldsymbol{\gamma}$ 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.

C.2 Mixing properties of ESSi

In this Section we report some stylised facts about the performance of the ESSi with τ fixed at 1. Figure S.3, 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 $\prod_{l=1}^{L}\left[p\left(\gamma_l\left|y\right.\right)\right]^{1/t_l}$, 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 S.3, bottom panels, shows the trace plot of the log posterior and the 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 S.3 which shows good overall acceptance rates for the collection of moves that we have implemented. The dimension of the problem does not seem to affect the acceptance 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 five 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 and the new block crossover, see Section 3.1 in the paper.

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 S.4, 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 $f(\gamma) = \log p(y|\gamma) + \log p(\gamma)$ across sweeps, confirm the good performance of ESSi. Again some instability is present in the high temperature chains, see in Table S.4 the overlapping index between chains $3, 4$ and $4, 5$ in Example 3 to 6.

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 S.3 about here – Table S.4 about here – Figure S.3 about here]

C.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:

$$p(\gamma_j = 1|y) \simeq C^{-1} \sum_{t=1,\dots,T} 1_{\left(\gamma_j^{(t)}=1\right)}(\gamma)\, p\left(y\left|\gamma^{(t)}\right.\right) p\left(\gamma^{(t)}\right) \tag{S.6}$$

with $C = \sum_{t=1,\dots,T} p\left(y\left|\gamma^{(t)}\right.\right) p\left(\gamma^{(t)}\right)$ and T the number of sweeps after the burn-in. The posterior model size is similarly defined, $p(p_\gamma|y) \simeq C^{-1} \sum_{t=1,\dots,T} 1_{\left(|\gamma^{(t)}|=p_\gamma\right)}(\gamma)\, p\left(y\left|\gamma^{(t)}\right.\right) p\left(\gamma^{(t)}\right)$, with C as before. Besides plotting the marginal posterior inclusion probability (S.6) averaged across sweeps

and replicates for our simulated examples, we will also compute the interquartile range of (S.6) 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 $p(\gamma|y)$ and R^2_γ : first we rank $p(\gamma|y)$ in decreasing order and record the indicator γ that corresponds to the maximum and $1,000$ largest $p(\gamma|y)$ (after burn-in). Given the above set of latent binary vectors, we then compute the corresponding R^2_γ leading to "R^2_γ: $\max p(\gamma|y)$" as well as the mean R^2_γ over the $1,000$ largest $p(\gamma|y)$, "$\overline{R^2_\gamma}$: $1,000$ largest $p(\gamma|y)$", both quantities averaged across replicates. Moreover the actual ability of the algorithm to reach regions of high posterior probability and persist on them is monitored: given the sequence of the $1,000$ best γs (based on $p(\gamma|y)$), the standard deviation of the corresponding R^2_γs 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 S.4 presents the marginal posterior probability of inclusion for ESSi with $\tau = 1$ averaged across replicates and, as a measure of variability, the interquartile 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 S.4 (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 S.4 (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 S.4 (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 S.4 are the median and interquartile range across replicates of $p(\gamma_j = 1 \,|y)$, $j = 1, \ldots, p$, 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, Figure S.4 (a), 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 $p(\gamma_j = 1 \,|y)$ (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 S.5. 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 $p(\gamma|y)$. 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 S.4. 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 $p + p_\gamma(p - p_\gamma)$, 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 S.5 about here – Figure S.4 about here]

C.4 Mixing properties and 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 S.5 we illustrate the performance of ESSg when the Z-S 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 S.5 (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 interquartile range shows non-negligible uncertainty about the estimates.

Figure S.5 (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 Z-S 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, 0.441 and 0.435 respectively.

The right panels show a complementary story. For ESSg, top right panel, $p(\tau \,|y\,)$, black solid line, leans quite far apart from the prior distribution, red solid line. The posterior mode is $3,953$, a value close but larger than the Benchmark prior proposed by Fernández *et al.* (2001) in the g-prior set-up showing that the calibration of τ *a priori* is complex and the choice between the Unit Information Prior, $\tau = n$, and Benchmark prior, $\tau = \max\left(n, p^2\right)$, is not straightforward. 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 1.496, a value quite far from 1 which is the recommended choice for τ after standardisation of the covariates. However in this particular example the influence of the hyperprior is negligible with respect to the fixed case, $\tau = 1$.

[Figure S.5 about here]

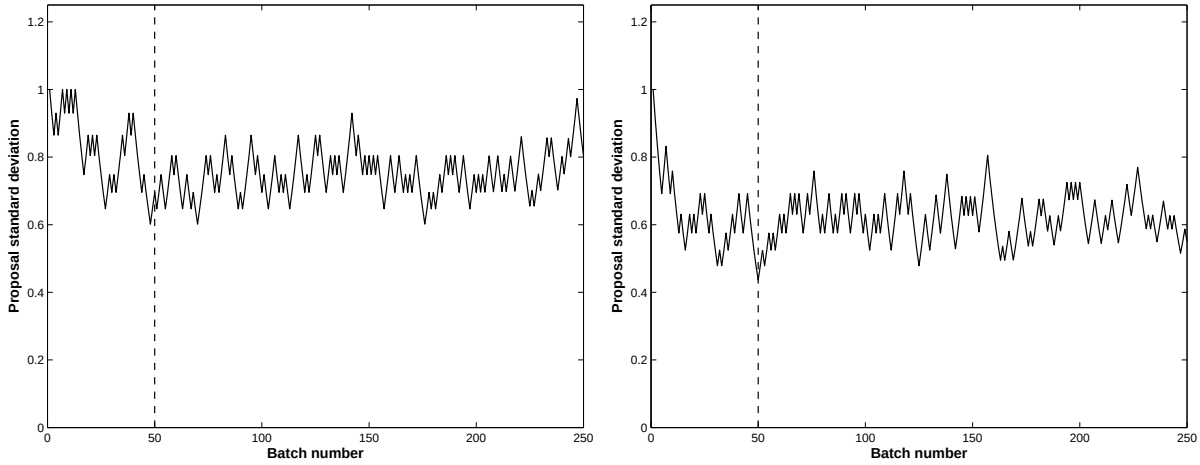
Further references

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions*. New York: Dover Publications, Inc.

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.

Iba, Y. (2001). Extended Ensemble Monte Carlo. *Int. J. Mod. Phys., C*, **12**, 623-656.

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



(a) (b)

Figure S.1: Trace plot of the proposal's standard deviation for τ for the two real data examples analysed using ESSg with Z-S prior. Vertical dashed lines indicate the end of the burn-in.

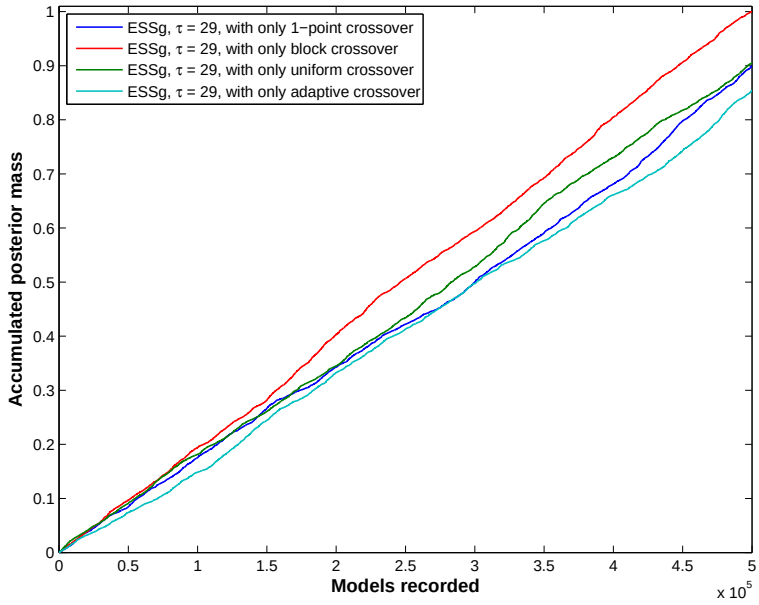


Figure S.2: Accumulated posterior mass as a function of the models recorded. Plot generated using 25 replicates of the analysis of the first real data example and normalised by the total mass found by ESSg, $\tau = 29$, with only block crossover move ($\rho_0 = 0.25$). 1-point and uniform crossover accumulate around 90% of the total mass accumulated by ESSg with only block crossover, while adaptive crossover only 85%.

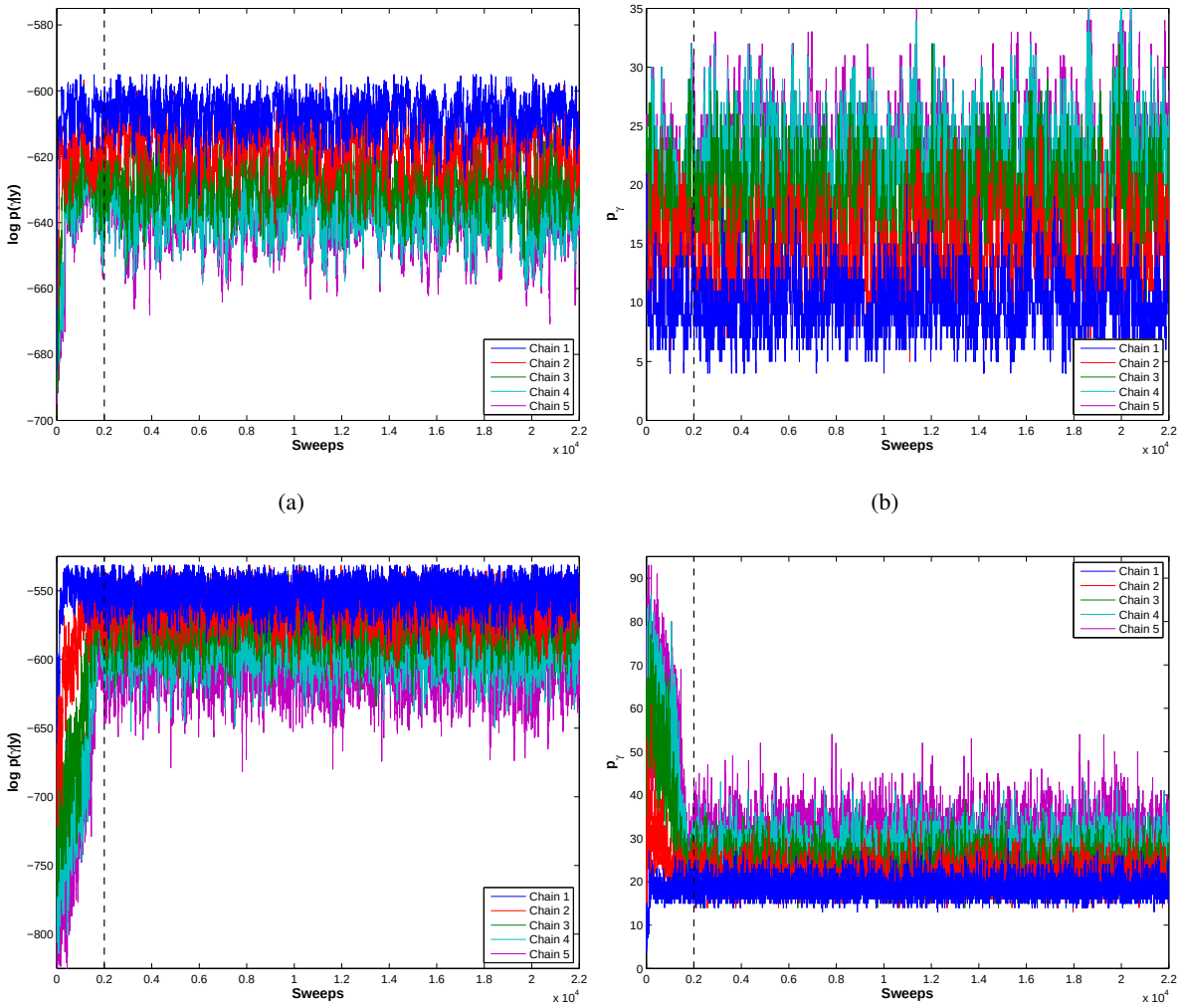


Figure S.3: For ESSi with $\tau = 1$: (a) trace plot of the log posterior probability, $\log p(\gamma | y)$, and (b) model size, p_γ, across sweeps for one replicate of Ex1 with $E(p_\gamma) = 20$, top panels and Ex4, bottom panels. Vertical dashed lines indicate the end of the burn-in.

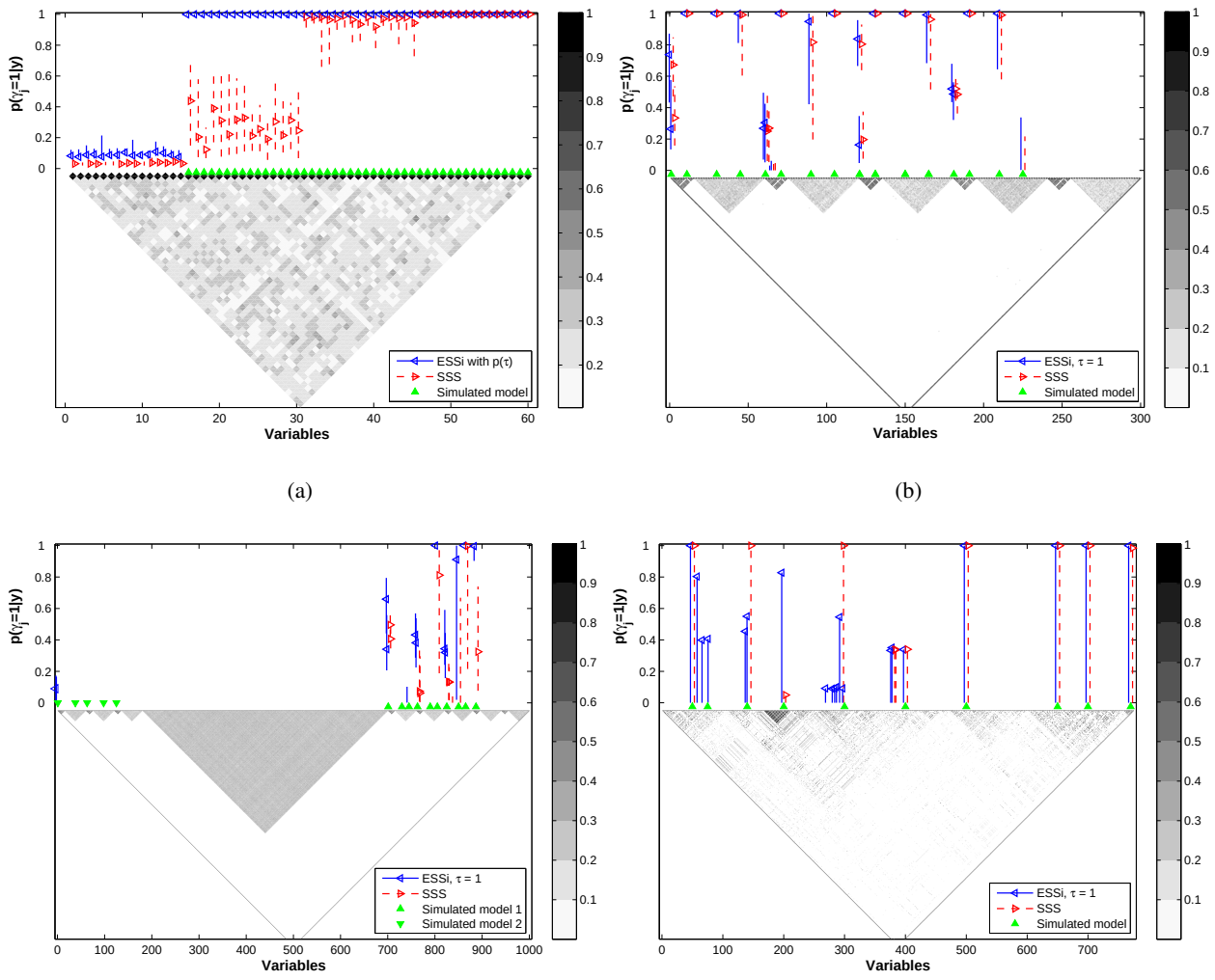
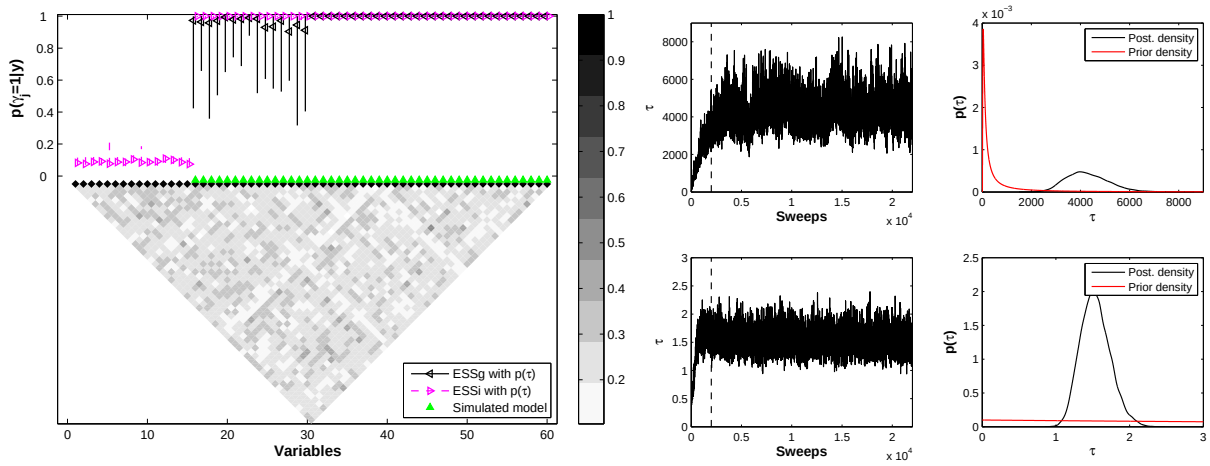


Figure S.4: Median and interquartile range of the marginal posterior probability of inclusion (S.6) for Ex3, (a), Ex4, (b) and Ex5, (c), across replicates. Each graph is constructed as follows: bottom part, pairwise squared correlation $\rho^2\left(X_j, X_{j'}\right)$, $j = 1, \ldots, p$, between predictors for one selected replicate, grey scale indicates different values of squared correlation; blue left and red right triangles, median of $p\left(\gamma_j = 1 \,|y\right)$ across replicates for ESSi with $\tau = 1$ and SSS respectively; vertical blue solid lines and vertical red dashed lines, interquartile range of $p\left(\gamma_j = 1 \,|y\right)$ 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 S.5: (a) Median and interquartile range of the marginal posterior probability of inclusion (S.6) across replicates for Ex3 when ESSg is applied with Z-S prior (median, back left triangles and interquartile range, vertical black solid lines) and ESSi is used with a proper but diffuse exponential prior centred in 10 (median, magenta right triangles and interquartile range, vertical magenta dashed lines). Upper green triangles, simulated models. Marginal posterior probability of inclusion lower than 0.025 not shown. Bottom part, pairwise squared correlation $\rho^2\left(X_j, X_{j'}\right)$, $j = 1, \ldots, p$, between predictors for one selected replicate, grey scale indicates different values of squared correlation. (b) Top panels, trace plot and posterior kernel density of τ for ESSg with Z-S 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.

	Version of ESSg	τ	$p(\tau)$
Experiment (i)	ESSg with only FSMH$_2$	68%	88%
	ESSg with only MC^3	28%	40%
Experiment (ii)	ESSg with only 1-point crossover	64%	80%
	ESSg with only block crossover	80%	84%
	ESSg with only uniform crossover	60%	84%
	ESSg with only adaptive	60%	76%

Table S.1: Proportion of times different versions of ESSg reach the same top visited model in the eQTL real data set with or without an hyperprior on τ in 25 replicates of the analysis.

	Version of ESSg	τ	$p(\tau)$
Experiment (ii)	ESSg with only 1-point crossover	0.303	0.335
	ESSg with only block crossover	0.482	0.501
	ESSg with only uniform crossover	0.026	0.042
	ESSg with only adaptive	0	0.013

Table S.2: Average acceptance rate of the crossover operator for different versions of ESSg in 25 replicates of the analysis of the first real data example with or without an hyperprior on τ.

	Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
$E(p_\gamma)$	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)
DR Exchange	0.500	0.493	0.500	0.582	0.492	0.517	0.505	0.497
	(0.040)	(0.043)	(0.040)	(0.020)	(0.071)	(0.105)	(0.013)	(0.072)

Table S.3: Mean and standard deviation in brackets of EMC acceptance rates across replicates for ESSi with $\tau = 1$. "DR Exchange" stands for "delayed rejection exchange".

		Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
$E(p_\gamma)$		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 S.4: Swapping probability for ESSi 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 $V(f(\gamma_l))(1/t_{l+1} - 1/t_l)^2$, Liang and Wong (2000) with $f(\gamma_l) = \log p(y|\gamma_l) + \log p(\gamma_l)$. Target value for consecutive chains is $O(1)$.

		Ex1			Ex2	Ex3	Ex4	Ex5	Ex6
	$E(p_\gamma)$	5	10	20	5	5	5	5	5
ESSi, τ=1	$R_\gamma^{2\,*}$	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)
	$\overline{R_\gamma^2}^{**}$	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 (min.)	6	6	7	16	18	166	338	202
		(< 1)	(< 1)	(< 1)	(< 1)	(1)	(32)	(43)	(40)
SSS	$R_\gamma^{2\,*}$	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)
	$\overline{R_\gamma^2}^{**}$	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 (min.)	12	12	13	118	497	502	169	549
		(1)	(2)	(2)	(26)	(75)	(241)	(81)	(159)

Table S.5: Comparison between ESSi with $\tau = 1$ and SSS for the six simulated examples. Standard deviation in brackets. $R_\gamma^{2\,*}$ and $\overline{R_\gamma^2}^{**}$ correspond to "R_γ^2: $\max p(\gamma|y)$" and "$\overline{R_\gamma^2}$: 1,000 largest $p(\gamma|y)$" respectively.