BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian models for sparse regression analysis of high dimensional data

Sylvia Richardson, Leonardo Bottolo & Jeffrey S. Rosenthal
Imperial College London, UK *University of Toronto, Canada*
sylvia.richardson@imperial.ac.uk l.bottolo@imperial.ac.uk jeff@math.toronto.edu

Summary

This paper considers the task of building efficient regression models for sparse multivariate analysis of high dimensional data sets, in particular it focuses on cases where the numbers q of responses $Y = (\underline{y}_k, 1 \leq k \leq q)$ and p of predictors $X = (\underline{x}_j, 1 \leq j \leq p)$ to analyse jointly are both large with respect to the sample size n, a challenging bi-directional task. The analysis of such data sets arise commonly in genetical genomics, with X linked to the DNA characteristics and Y corresponding to measurements of fundamental biological processes such as transcription, protein or metabolite production. Building on the Bayesian variable selection set-up for the linear model and associated efficient MCMC algorithms developed for single responses, we discuss the generic framework of hierarchical related sparse regressions, where parallel regressions of $\underline{y}_k$ on the set of covariates X are linked in a hierarchical fashion, in particular through the prior model of the variable selection indicators γ_{kj}, which indicate among the covariates $\underline{x}_j$ those which are associated to the response $\underline{y}_k$ in each multivariate regression. Structures for the joint model of the γ_{kj}, which correspond to different compromises between the aims of controlling sparsity and that of enhancing the detection of predictors that are associated with many responses ('hot spots'), will be discussed and a new multiplicative model for the probability structure of the γ_{kj} will be presented. To perform inference for these models in high dimensional set-ups, novel adaptive MCMC algorithms are needed. As sparsity is paramount and most of the associations expected to be zero, new algorithms that progressively focus on part of the space where the most interesting associations occur are of great interest. We shall discuss their formulation and theoretical properties, and demonstrate their use on simulated and real data from genomics.

Keywords and Phrases: Adaptive MCMC scanning, eQTL, Genomics, hierarchically related regressions, variable selection.

SR and LB contributed equally to this work. The support of MRC grant G0600609 is gratefully acknowledged. LB acknowledges the support of MRC Clinical Sciences Center. We thank Krzysztof Łatuszyński for very helpful comments.

1. INTRODUCTION

The size and diversity of newly available genetic, genomics and other 'omics data sets has meant that going beyond the finding of strong univariate (or low dimension) associations to reveal more complex patterns related to the underlying biological pathways and metabolism has proved difficult. The current focus of much biological research has now moved to Integrative Genomics, which encompasses a variety of biological questions involving the *combined analysis* of any two or more types of genomics data sets. For example investigations into the genetic regulation of transcription or metabolite synthesis, so called eQTL or mQTL studies, or into the influence of copy number variations on expression are carried out to progress understanding of the function of genes. Research on how to *jointly model* two or more such highly dimensional data sets, with different intrinsic structures and scale of measurements, is thus a key priority and a difficult challenge for statisticians. The Bayesian modelling paradigm is particularly well suited to address complex questions regarding structural links between different pieces of data, for building in hierarchical relationships based on substantive knowledge, for adopting prior specifications that translate expected sparsity of the underlying biology and for uncovering a range of alternative explanations. On the other hand, the computational challenges faced by any joint analysis of high dimensional data are substantial, resulting in relatively few fully Bayesian analyses being attempted.

In this paper, we propose to carry out sparse multivariate analysis of high dimensional data sets by developing a framework of hierarchically related sparse regressions to model the association between large numbers of responses (e.g. measurements of gene expression), $Y = (\underline{y}_1, \ldots, \underline{y}_k, \ldots, \underline{y}_q)$, $\underline{y}_k = (y_{1k}, \ldots, y_{ik}, \ldots, y_{nk})^T$ recorded on n subjects, and a large number of predictors (e.g. a set of discrete genetic markers for each subject), recorded in the form of a matrix $n \times p$ ($n \ll p$) of covariates $X = (\underline{x}_1, \ldots, \underline{x}_j, \ldots, \underline{x}_p)$, $\underline{x}_j = \left(x_{1j}, \ldots, x_{ij}, \ldots, x_{nj}\right)^T$). A fully multivariate model that would treat *all* the responses as a vector and link its distribution to *all* the predictors is neither feasible when p and q are both in their thousands, nor appropriate as the biological context suggests that we should expect sparse associations between each response and the predictors. Of major interest is the existence of so called 'hot spots', i.e. finding genetic markers $\underline{x}_j$ that show evidence of enhanced linkage, i.e that are associated to many responses, as this indicates that this region of the genome might play a key regulatory role. To tease out such structure, we propose to model the relationship between Y and X in a hierarchical fashion, first associating each response with a small subset of the predictors via a subset selection formulation, and then linking the selection indicators in a hierarchical manner. We show that by empowering MCMC algorithms with features such as parallel tempering/evolutionary Monte Carlo and adaptive schemes, we can make such models workable for realistic joint analyses in genomics. In particular, we propose a new class of adaptive scanning schemes, give conditions that ensure their theoretical properties and highlight their benefits on simulated data sets and an eQTL experiment from a study of diabetes in mice.

2. BAYESIAN MODELS IN GENETICAL GENOMICS

Much of the recent work on joint analysis of high dimensional data has been motivated by the framework of eQTL studies (expression Quantitative Trait Loci) where the responses are quantitative measures of gene expression abundances for a thousands of transcripts and the predictors encode DNA sequence variation at a large

number of loci. In turn, eQTL analyses have built upon models for multiple mapping of Quantitative Trait Loci (QTL), also referred to as polygenic models, i.e. models where the aim is to quantify the association of a single continuous response, referred to as a 'trait', with DNA pattern at multiple genetic loci by using a sparse multivariate regression approach.

2.1. *Bayesian multiple mapping for Quantitative Trait*

It is not our purpose to discuss comprehensively the work on Bayesian multiple mapping for quantitative trait, see Yi and Shriner (2008) for a recent review. As expected, several styles of approaches to variable selection have been taken, differing principally in the choice of priors for the regression coefficients linking the trait with the genetic markers and in the adopted prior specification of the model space.

Most commonly, QTL studies have adopted a Bayesian variable selection formulation which starts from the full linear model and considers independent priors for the regression coefficients β_j, introducing variable selection via auxiliary indicators $\gamma_j, 1 \leq j \leq p$ where $\gamma_j = 1$ encodes the presence of the j^{th} covariate in the linear model. As reviewed by O'Hara and Sillanpää (2009), such implementations differ in the way the joint prior for (γ_j, β_j) is defined. Independent priors for γ_j and β_j proposed by Kuo and Mallick (1998) have been used in Bayesian mapping but sometimes lead to instability (O'Hara and Sillanpää, 2009). In most other works, a decomposition $p(\beta_j, \gamma_j) = p(\beta_j \mid \gamma_j)p(\gamma_j)$ is used, leading to independent mixture priors for each β_j in the form of a spike component at or around zero and a flat slab elsewhere, inspired by the stochastic search variable selection (SSVS) approach proposed by George and McCullogh (1993).

Note that specifying priors for the regression parameters of the full linear model may be inappropriate when the regressors are not orthogonal as the coefficients have a different interpretation under submodels corresponding to different $\underline{\gamma}$ vectors (Ntzoufras, 1999). An alternative formulation that defines priors for regression coefficient *conditional on the whole vector* $\underline{\gamma}$ might be preferable. Moreover, such a formulation allows the regression coefficients to be integrated out, facilitating the implementation of algorithms that sample the model space of the selection indicators, referred to as subset selection algorithms (Clyde and George, 2004). Such specification naturally leads to the so-called g-prior formulation which encodes a correlation structure between the regression coefficients that reproduces the covariance structure of the likelihood. In genetic applications, this would appear most appropriate in view of the complex structure of the X induced by population structure.

2.2. *Bayesian eQTL models*

The framework of eQTL experiments is aimed at understanding the genetic basis of regulation by (i) treating the high dimensional set of gene expression as multiple responses and (ii) uncovering their association with the genetic markers. Markers with evidence of enhanced linkage, 'hot spots' are of particular interest. The first analyses were carried out by repeated application of simple univariate QTL analyses for each transcript, without attempting to share any information across transcripts or to account for multiple mapping.

The first joint approach which aimed at modelling all the transcripts via a mixture formulation was proposed by Kendziorski *et al.* (2006). In the Mixture Over Markers (MOM) approach, each response (expression value of a transcript) $\underline{y}_k, 1 \leq k \leq q$, is linked to the marker j with probability p_j and assumed to then fol-

low a distribution $f_j(\cdot)$ common for all the transcripts mapping to marker j. In complement, with probability p_0, a response is not linked to any marker, and those non-mapping transcripts have distribution $f_0(\cdot)$. The marginal distribution of the data for each response $\underline{y}_k$ is thus given by a mixture model: $p_0 f_0(\underline{y}_k) + \sum_{j=1}^{p} p_j f_j(\underline{y}_k)$. A basic assumption of this model is that a response is associated with *at most one predictor* (genetic marker). Information from all the responses associated to a particular marker j is then used to estimate $f_j(\cdot)$. For good identifiability of the mixture, MOM requires a sufficient number of transcripts to be associated with the markers. Using thresholds on the posterior probabilities p_j based on preset false discovery rate (FDR) control, each response can be associated with the most likely location (or no location at all) and the fraction of responses associated with each marker j can be used to detect hot spots. By combining information across the responses, MOM has a better control of FDR than pure univariate methods.

Noting that the formulation of Kendziorski *et al.* (2006) is limited to monogenic mapping, Jia and Xu (J & X) (2007) set up the search for eQTL associations into a single model where each transcript $\underline{y}_k, 1 \leq k \leq q$ is potentially linked to the full set of p markers X through a full linear model with regression coefficients, $\underline{\beta}_k = (\beta_{k1}, \ldots, \beta_{kj}, \ldots, \beta_{kp})^T$. Inspired by QTL models and SSVS variable selection, they use a mixture prior on each of the β_{kj}:

$$\beta_{kj} \sim (1 - \gamma_{kj}) N(0, \delta) + \gamma_{kj} N(0, \tau_k^2)$$

with a fixed very small δ for the spike and a hierarchical prior for the variances τ_k^2 of the slabs. They then link the q responses through a model of the indicators γ_{kj}, $\gamma_{kj} \sim Bernoulli(\zeta_j)$, establishing what we refer to as a *hierarchical regression set-up*. J & X linked regression set-up shares common features with ours. We will discuss this further in Section 3.2 and present a comparison of their algorithm BAYES with ours on simulated data in Section 6.3.

A third class of models for joint analysis is that of stochastic partition models for association. This approach, proposed by Monni and Tadesse (M & T) (2009), partitions the responses into disjoint subsets or clusters that have a *similar dependence* on a subset of covariates (or no dependence). M & T implement such a model for analysing the association between genomic CGH data and gene expression. In their set-up, each response cluster C is associated with a subset of response indices, $Q(C)$ and a subset of predictor indices $P(C)$, in such a way that all the $\underline{y}_k$ in cluster C are linked to *the same subset of predictors via the same regression coefficients*: $\beta_{kj} = \beta_j,\ k \in Q(C), j \in P(C)$. This assumption on the βs may be appropriate in some context, but is quite restrictive in general. M & T allow for response specific intercept and cluster specific noise. Dimension reduction and borrowing of information is obtained through the sharing of a common parameter in the cluster. Their prior formulation for the regression coefficients is conditional on the cluster and they exploit conjugacy to integrate these out in order to improve mixing. They assign product priors to configurations that penalise large clusters through a tuning parameter ρ and use reversible jump moves and parallel tempering to search through the high dimensional space of partitions, acknowledging that such a search is challenging. In their output, they mostly consider the MAP (maximum a posteriori) configuration.

3. MODELLING $Y|X$: HIERARCHICAL RELATED SPARSE REGRESSION

In order to discover the pattern of association between subgroups of Ys and the predictors, we model the relationship between Y and X through q regression equations linked by a hierarchical model on the variable selection process.

3.1. *Subset selection*

We define the q regression equations as $\underline{y}_k = \alpha_k 1_n + X\underline{\beta}_k + \underline{\epsilon}_k$, $k = 1, \ldots, q$, where $\underline{\epsilon}_k \sim N_n\left(\underline{0}, \sigma_k^2 I_n\right)$. Note that every regression equation has its own intercept α_k and error variance σ_k^2. In order to perform variable selection, i.e. to find a sparse subset of predictors that explain the variability of Y but there is uncertainty about which subset to use, we introduce a latent binary vector $\underline{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kj}, \ldots, \gamma_{kp})^T$ for each regression equation where $\gamma_{kj} = 1$ if $\beta_{kj} \neq 0$ and $\gamma_{kj} = 0$ if $\beta_{kj} = 0$, $j = 1, \ldots, p$. Considering all the q regressions, we obtain the $q \times p$ latent binary matrix $\Gamma = \left(\underline{\gamma}_1, \ldots, \underline{\gamma}_k, \ldots, \underline{\gamma}_q\right)^T$. Adopting the subset selection formulation and assuming independence of the q regression, given Γ, the likelihood becomes

$$\prod_{k=1}^{q} \left(\frac{1}{2\pi\sigma_k^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma_k^2}\left(\underline{y}_k - \alpha_k 1_n - X_{\gamma_k}\underline{\beta}_{\gamma_k}\right)^T \left(\underline{y}_k - \alpha_k 1_n - X_{\gamma_k}\underline{\beta}_{\gamma_k}\right)\right\}, \quad (1)$$

where $\underline{\beta}_{\gamma_k}$ is the non-zero vector of regression coefficients of the k^{th} regression and X_{γ_k} is the design matrix with columns corresponding to $\gamma_{kj} = 1$.

3.2. *Priors*

As discussed in Section 2.2 we follow a g-priors representation for the regression coefficients. Conditionally on $\underline{\gamma}_k$, we assume:

$$\underline{\beta}_{\gamma_k} \,|\, \underline{\gamma}_k, g, \sigma_k^2 \sim N_{p_{\gamma_k}}\left(\underline{0}, g\left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1}\sigma_k^2\right), \quad (2)$$

where $p_{\gamma_k} \equiv \underline{\gamma}_k^T 1_p$ is the number of non-zero elements in $\underline{\gamma}_k$. To increase flexibility, the level of shrinkage g is not fixed but given a hyperprior: $g \sim InvGam\,(a_g, b_g)$. Note however that the level of shrinkage is *common* for all the q regression equations, so g is one of the parameters that links the q regressions.

Prior specification is completed by assigning a Bernoulli prior on the latent binary indicators:

$$p\left(\gamma_{kj}\,|\omega_{kj}\right) = \omega_{kj}^{\gamma_{kj}}\left(1-\omega_{kj}\right)^{1-\gamma_{kj}}, k = 1, \ldots, q, j = 1, \ldots, p.$$

Modelling the matrix of the prior probabilities for Γ

$$\Omega = \begin{bmatrix} \omega_{11} & \cdots & \omega_{1j} & \cdots & \omega_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{k1} & \cdots & \omega_{kj} & \cdots & \omega_{kp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{q1} & \cdots & \omega_{qj} & \cdots & \omega_{qp} \end{bmatrix}$$

is crucial as this is where considerations of sparsity and borrowing of strength between the responses can be included. Three strategies can be adopted:

(i) $\omega_{kj} = \omega_k$ with $\omega_k \sim Beta\left(a_{\omega_k}, b_{\omega_k}\right)$;

(ii) $\omega_{kj} = \omega_j$ with $\omega_j \sim Beta\left(c_{\omega_j}, d_{\omega_j}\right)$;

(iii) $\omega_{kj} = \omega_k \times \rho_j$ with $\omega_k \sim Beta\left(a_{\omega_k}, b_{\omega_k}\right)$, $\rho_j \sim Gam\left(c_{\rho_j}, d_{\rho_j}\right)$, $0 \leq \omega_{kj} \leq 1$;

We refer to model (i) as 'the independent model', to model (ii) as 'the column effect model' and finally we name model (iii) 'the multiplicative model'. The first model assumes that the underlying selection probabilities for each response $\underline{y}_k$ may be different and arise from independent Beta distributions. It is a direct extension of the variable selection model for single response in Bottolo and Richardson (2010). The only shared parameter among the q responses is the shrinkage coefficient g. Linking the k regressions through g is natural in view of the similarity of the responses in our set-up and helps to stabilise the effect of g. The second model is inspired by Jia and Xu (2007) and introduce a shared parameter ω_j (which plays a similar role to their parameter ζ_j) which quantifies the probability for each predictor to be associated with any, possibly many, transcripts. In a simplistic manner, model (ii) assumes that this probability is the same for all the responses. Finally the third model is a new extension of the previous two. A shared column effect ρ_j is used to moderate the underlying selection probability ω_k specific to the k^{th} regression in a multiplicative fashion, which combines the good features of models (i) and (ii).

Models (i) and (iii) share an important feature: the hyper parameters a_{ω_k} and b_{ω_k} can be easily related to an elicited prior mean and variance for p_{γ_k}, the number of predictors. Context specific knowledge on the expected sparsity of the regressions, e.g. information on a typical range for the number of genetic associations, can thus inform choices for a_{ω_k} and b_{ω_k}. Note also that in model (i), it is possible to integrate out ω_k, while in model (ii) and (iii) ω_j and (ω_k, ρ_j) will need to be sampled.

The most important difference between the models we are considering is the way sparsity can – or cannot – be induced. In contrast to models (i) and (iii), in model (ii) the simple column structure has destroyed any possible control on the expected number of associations. The ω_j in model (ii) are directly related to the relative proportion of the q outcomes that are associated with the j^{th} covariate, and will be hardly influenced by choices of c_{ω_j} and d_{ω_j} values. It will be interesting to see how this formulation of shared column effect within a subset selection approach performs in comparison to the column model of J & X with SSVS variable selection, and whether there are any problems of over-estimation of hot spots.

Model (iii) synthesizes the benefits of models (i) and (ii): for each response the level of sparsity can be informed through the hyper parameters a_{ω_k} and b_{ω_k} while ρ_j captures the 'propensity' for predictor j to influence several outcomes at the same time. In this model, the role of ρ_j can be seen as a 'predictor specific propensity for being a hot spot' that inflates/deflates the underlying selection level ω_k. The adopted multiplicative formulation has some similarity to the disease mapping paradigm where relative risks act in a multiplicative fashion on expected number of cases in a binomial or Poisson disease risk model. Accordingly, we decided to center ρ_j on 1 and choose $c_{\rho_j} = d_{\rho_j} = 1.2$ so that the coefficient of variation is reasonably large, but that there is not much probability mass on small values. The benefit of having split ω_{kj} into two components, which will be given independent priors, is that we have allowed borrowing of information across the responses without destroying the possibility of inducing sparsity. In the following, we will focus our investigations

on models (ii) and (iii). As model (i) does not borrow any information across the responses, it is less adapted to tease out 'hot spot' structure.

We end this section by discussing the hyper parameters for the priors of g and σ_k^2. In this paper we use the values proposed in Bottolo and Richardson (2010), with $a_g = 1/2$ and $b_g = n/2$, so that (2) can be thought as a mixture of g-priors and an inverse-gamma prior with non-existing moments. Finally we specify a relative flat prior for the error variance selecting $a_\sigma = 10^{-3}$ and $b_\sigma = 10^{-6}$.

Given the likelihood, the prior structure, natural conditional independence assumptions, and after integrating out the intercepts, the regression coefficients, and the error variances, the joint density can be written as

$$p(g) \prod_{k=1}^{q} p\left(\underline{y}_k \,|\, X, \underline{\gamma}_k, g\right) p\left(\underline{\gamma}_k \,|\, \underline{\omega}_k\right) p(\underline{\omega}_k), \tag{3}$$

where $\underline{\omega}_k = (\omega_{k1}, \ldots, \omega_{kj}, \ldots, \omega_{kp})^T$, with the likelihood for the k^{th} regression given by

$$p\left(\underline{y}_k \,|\, X, \underline{\gamma}_k, g\right) \propto (1+g)^{-p_{\gamma_k}/2} \left(2b_\sigma + S\left(\underline{\gamma}_k\right)\right)^{-(2a_\sigma + n - 1)/2} \tag{4}$$

with $S\left(\underline{\gamma}_k\right) = \left(\underline{y}_k - \bar{\underline{y}}_k\right)^T \left(\underline{y}_k - \bar{\underline{y}}_k\right) - \frac{g}{1+g} \left(\underline{y}_k - \bar{\underline{y}}_k\right)^T X_{\gamma_k} \left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1} X_{\gamma_k}^T \left(\underline{y}_k - \bar{\underline{y}}_k\right)$ and $\bar{\underline{y}}_k = 1_n \sum_{i=1}^{n} y_{ik}/n$.

4. MCMC ALGORITHM

The task of updating all variables in this *large* $p \times$ *large* q set-up is very demanding computationally. To do this, we have assembled key ingredients – parallel tempering/evolutionary Monte Carlo and adaptive moves – in a new algorithm, Hierarchical Evolutionary Stochastic Search, HESS hereafter.

For each of the q regressions, we consider L chains, with temperature t_{kl}, $1 = t_{k1} < t_{k2} < \cdots < t_{kL}$, where t_{kl} is the temperature attached to the l^{th} chain in the k^{th} regression. $L = 1$ corresponds to the non-heated chain, and only variables in the non-heated chain are retained in the final output of the algorithm. We denote by $\underline{\gamma}_{kl} = (\gamma_{kjl}, 1 \le j \le p)$ and $\underline{\omega}_{kl} = (\omega_{kjl}, 1 \le j \le p)$ the vectors of selection indicators and probabilities respectively for the l^{th} chain of the k^{th} regression. The variables that will be updated during a sweep of HESS are in turn $\left(\{\underline{\gamma}_{kl}\}, \{\underline{\omega}_{kl}\}, 1 \le k \le q, 1 \le l \le L\right)$ and g. The following full conditionals will be used throughout in the relevant acceptance ratios

$$p\left(\underline{\gamma}_{kl} \,|\, \cdots\right)^{1/t_{kl}} \propto p\left(\underline{y}_k \,|\, X, \underline{\gamma}_{kl}, g\right)^{1/t_{kl}} p\left(\underline{\gamma}_{kl} \,|\, \underline{\omega}_{kl}\right)^{1/t_{kl}}, \tag{5}$$

$$p\left(\underline{\omega}_{kl} \,|\, \cdots\right)^{1/t_{kl}} \propto p\left(\underline{\gamma}_{kl} \,|\, \underline{\omega}_{kl}\right)^{1/t_{kl}} p\left(\underline{\omega}_{kl}\right)^{1/t_{kl}}, \tag{6}$$

$$p\left(g \,|\, \cdots\right) \propto p(g) \prod_{l=1}^{L} \prod_{k=1}^{q} p\left(\underline{y}_k \,|\, X, \underline{\gamma}_{kl}, g\right)^{1/t_{kl}}. \tag{7}$$

The update of the packet $\left(\{\underline{\gamma}_{kl}\}, 1 \le k \le q, 1 \le l \le L\right)$ builds on the Evolutionary Stochastic Search (ESS) algorithm of Bottolo and Richardson (2010) and is briefly

described in Section 4.1. For the update of the matrix Ω of joint selection probabilities, we have used an adaptive sampler described in Section 4.2. The scanning strategy which features a novel scheme of adaptive scanning over k is discussed in Section 4.3.

4.1. *Recall of main ESS scheme for the* Γ *updates*

The key features of ESS that we exploit here is the use of evolutionary Monte Carlo (EMC) to explore the huge model space as well as an automatic tuning of the temperature placement during burn-in. Multiple chains are run in parallel at different 'temperatures' with two distinct type of moves: (i) local moves aimed at updating the indicators of every single chain and (ii) global moves (crossover and exchange operators) that try to exchange part or the whole configuration of $\underline{\gamma}_{kl}$ for selected chains. Global moves are important because they allow the algorithm to escape from local modes, while a detailed exploration is left to the local moves. While global moves are computationally inexpensive, the local ones could be time costing (e.g. full Gibbs sampling over j is prohibitive). In ESS, a fast-scan Metropolis-within-Gibbs scheme for updating a set of γ_{kjl} was proposed, which includes an additional probability step to choose the indices where to perform the Metropolis-with-Gibbs update based on current model size and temperature. Here, we adopt a similar idea, but modify this additional step to use the current values of ω_{kjl} (which are available in our HESS set-up but were integrated out in ESS).

In summary, we carry out the update of $\left(\{\underline{\gamma}_{kl}\}, 1 \leq k \leq q, 1 \leq l \leq L\right)$ using the portfolio of global and local moves described in Bottolo and Richardson (2010) with obvious modifications to include the $\underline{\omega}_{kl}$ in the acceptance rates, following (5).

4.2. *g and* Ω *updates in HESS*

The variable selection coefficient g is common to *all* the q regression equations and to *all* L chains, see (7). The MCMC update of g is not particularly difficult and we implement a simple Metropolis-within-Gibbs with lognormal proposal density. For improving the mixing, we update g frequently.

For simplicity of notation, for the rest of this section, we shall not index variables by the chain index l, but stress that the description below applies to each chain. The update of Ω depends on whether model (ii) or model (iii) are considered as prior structure for Ω. Recall that in model (i), Ω is integrated out.

In model (iii), $\omega_{kj} = \omega_k \times \rho_j$. In this case, we found useful to update the scalars ω_k and ρ_j using a Metropolis-within-Gibbs sampler, based on (6), with random walk proposals and adaptive standard deviations, following Roberts and Rosenthal (2009). We use fixed non-overlapping batch of say, 50 sweeps, indexed by m. Denoting by $s_k(m)$ and $s_j(m)$ the proposal standard deviations at the m^{th} batch for updating ω_k and ρ_j respectively, we use random walk Metropolis and propose new values for ω_k and ρ_j: $\text{logit}(\omega_k') \sim N\left(\text{logit}(\omega_k), s_k^2(m)\right)$ and $\log\left(\rho_j'\right) \sim N\left(\log(\rho_j), s_j^2(m)\right)$. During the batch we monitor the acceptance rate, and use the adaptive update: $s_k(m+1) = s_k(m) \pm \delta_k(m)$, to guide the acceptance rate towards 0.44, and proceed similarly for the update of $s_j(m)$. We further impose the following restrictions in order to satisfy the conditions in Section 5:

$$\forall k, \quad M_{\omega 1} < s_k(m) < M_{\omega 2} \tag{8}$$

$$\forall k, \quad \delta_k(m) = \min\{\delta_\omega, m^{-1/2}\} \tag{9}$$

for some finite $M_{\omega 1}$ and $M_{\omega 2}$, and some $\delta_\omega > 0$, and impose similar restrictions for $s_j(m)$.

In model (ii), $\omega_{kj} = \omega_j$. In this case for the non-heated chain, the full conditional for ω_j is available in closed form. For the heated chains, we use again an Metropolis-within-Gibbs with adaptive proposals, similar to that for model (iii).

4.3. *Scanning strategy for updating the responses*

We now describe one of the distinguishing features of the HESS algorithm, the strategy for selecting the indices of the responses k to be updated. As q is large, it is important to investigate scanning strategies over k that can make use of potential sparsity in the q direction. The simplest one is to choose to update only a fraction $\phi, 0 < \phi < 1$ of the q responses at every sweep, i.e. to choose at random without replacement a group of responses of size $\phi \times q$ to update. We shall refer to this strategy as '*scanning with fixed fraction* ϕ'. In the eQTL context, only a moderate proportion of the gene expressions are expected to be under genetic control, and so it seems reasonable to update a fraction of, say, $\phi = 0.25$ of responses at every sweep (different fractions can be used if so required, informed by the expected percentage of responses a priori linked to any predictor).

An obvious limitation of the fixed ϕ scanning is that by choosing purely at random the fraction of the responses to update, we will end up updating many 'uninteresting' responses, i.e responses which are not associated with any predictor. It is thus of particular interest to investigate *new adaptive scanning strategies* which can learn the 'interesting' responses as the algorithm proceeds and progressively incorporate this knowledge into the scanning probabilities. In other words, we want to increase the probability of updating the selection indicators $\underline{\gamma}_{kl}$ for a response $\underline{y}_k$ *when this response is likely associated with several predictors*. Indeed, we know that variable selection is hard when p is large and therefore accomplishing more updates for these 'active' responses should improve the performance of the algorithm. To the best of our knowledge, such adaptive scanning strategies (which are different from adaptive random scans) have not been studied before and we refer to Section 5 for their theoretical properties. Here, we give details of the strategies that we have explored.

4.3.1. *Adaptive scanning*

We construct a vector $w_k = w_k(b)$, of selection probabilities, $k = 1, \ldots, q$ that will evolve as the algorithm progresses, where b increments a batch index, say, every 50 sweeps. We begin with a new definition of 'batch'. Differently from the fixed disjoint batches used in the Ω update, here the definition of batch must fulfill two conditions (to satisfy (C7) in Section 5): (i) the size (number of sweeps) of the b^{th} batch must grow to infinity, and (ii) two consecutive batches must share part of the chain history, such that the fraction of the two batches which overlaps converges to 1 as the algorithm proceeds.

These two conditions can be guaranteed in several ways. The simplest is to use a 'full memory batch size growth' at every S sweeps, say 50, the batch $(b+1)$ is defined as the complete chain history from the initial sweep. What we have implemented is a different batch definition where the influence of the start of the algorithm is progressively discarded: we use growing batches of size $bS - \left\lfloor \sqrt{bS} \right\rfloor$ ($\lfloor \cdot \rfloor$ integer part), so that the initial $\left\lfloor \sqrt{bS} \right\rfloor$ sweeps are removed from the history.

In each batch we monitor $\tilde{r}_k(b) = \sum_{s=1}^{S(b)} p_{\gamma_k}^{(s)} / S(b)$, where $p_{\gamma_k}^{(s)}$ is be the number of predictors included in the model for the k^{th} regression at sweep s and $S(b)$ is the total number of sweeps in the bth batch (we omit the l index since the selection probabilities are based on the non-heated chain, $l = 1$).

Next we introduce a function of the parameters that we will use to characterise the 'interesting' responses. Whereas different parameters can be monitored, the idea of tracking those responses with large $\tilde{r}(b)$ (and large $p_{\gamma_k}^{(s)}$ on average) is appealing as discussed previously.

Adaptive scanning scheme.

(i) At the end of each batch, we monitor $r(b)$, the renormalised version of $\tilde{r}_k(b)$ across the q responses;

(ii) To satisfy the theoretical conditions of Section 5, we set

$$\tilde{w}_k(b) = (1 - \varepsilon(b)) r_k(b) + \varepsilon(b) \tag{10}$$

for some $\varepsilon(b) > 0$. At the beginning and for a fixed number $b_0 S$ of sweeps, we let the algorithm explore all the responses with equal probability ($\varepsilon(b) = 1$). There is no adaptation of scanning probabilities during this period, and the algorithm uses the fixed fraction ϕ version. During this burn-in period, the algorithm accumulates an increasing quantity of 'memory' that will be used afterwards to derive good selection probabilities. After the burn-in stage, $\varepsilon(b)$ starts to decrease $\propto 1/b$ rate

$$\varepsilon(b) = \begin{cases} 1 & \text{if } b \leq b_0 \\ \frac{1}{c}\frac{b_0}{b} + 10^{-3} & \text{if } b > b_0 \end{cases} \tag{11}$$

where $c > 1$ is a constant that can be used to accelerate the decay of $\varepsilon(b)$;

(iii) We obtain the selection probabilities $w_k(b)$ renormalising $\tilde{w}_k(b)$ across the q responses.

(iv) Finally the vector of selection probabilities $w_k(b)$ are used to select at random without replacement a fraction ϕ of responses to be updated.

As will be explained in Section 5, if an adaptive scanning strategy is used, additional conditions on all the variables updated and the kernels must be imposed. To guarantee these conditions, we further impose (by rejecting any proposed move which violates any of the following constraints) that, for some $\eta > 0$ (depending of the model selected (ii) or (iii) for Ω):

$$\eta \leq g \leq 10^{10},\ \eta \leq \omega_{jl} \leq 1 - \eta,\ \eta \leq \omega_{kl} \leq 1 - \eta,\ \text{and } \eta \leq \rho_{jl} \leq 10^{10}. \tag{12}$$

5. THEORETICAL JUSTIFICATION

For ordinary MCMC algorithms, it is well known that basic properties such as ϕ-irreducibility and aperiodicity suffice to guarantee ergodicity (i.e., asymptotic convergence to the stationary distribution). However, some of the algorithms considered in this paper are *adaptive*, i.e. the transition probabilities change over time and may depend upon the chain's previous history. Such adaptions can easily destroy ergodicity, and it is known (see e.g. Andrieu and Moulines, 2006; Roberts and Rosenthal, 2007, 2009; and references therein) that use of adaptive algorithms requires careful theoretical justification.

For notation, let $\pi(\cdot)$ be the target density on the state space $\mathcal{X}$, let $U_n \in \mathcal{X}$ be a vector representing the full state of the adaptive algorithm at time n (including the $\underline{\gamma}_k$, g, ω_k, ρ_j, etc.; thus, $\mathcal{X}$ is part discrete and part continuous), and let $V_n \in \mathcal{Y}$ be a

vector representing all the associated adaptive parameters at time n (including the $s_k(m)$, $s_j(m)$, $\tilde{r}_k(b)$, $\varepsilon(b)$, etc.). For each fixed $v \in \mathcal{Y}$, let $P_v(u, \cdot)$ be the non-adaptive Markov chain kernel corresponding to that fixed choice of adaptive parameters, so

$$\mathbf{P}[U_{n+1} \in B \,|\, U_n = u, V_n = v, U_{n-1}, \ldots, U_0, V_{n-1}, \ldots, V_0] \;=\; P_v(u, B)$$

for all $u \in \mathcal{X},\ v \in \mathcal{Y},\ B \subseteq \mathcal{X}$, while the conditional distribution of V_{n+1} given the past is specified by the adaptive algorithm. We require the following conditions.

(C0) For all $u \in \mathcal{X}$ and each fixed $v \in \mathcal{Y}$, $\lim_{n\to\infty} \|P_v^n(u, \cdot) - \pi(\cdot)\| = 0$, where $\|P_v^n(u, \cdot) - \pi(\cdot)\| = \sup_{B\subseteq\mathcal{X}} |P_\gamma^n(u, B) - \pi(B)|$ is total variation distance.

(C1) The subsets $\mathcal{X}$ and $\mathcal{Y}$ are both compact.

(C2) There is a finite collection $\mathcal{S}$ of sequences of coordinates, such that each kernel P_v is defined by first selecting a sequence $s \in \mathcal{S}$ according to some selection probabilities $p_v(s)$, and then applying successive Metropolis-Hastings-within-Gibbs iterations (possibly adaptive or possibly pure Gibbs) to each variable in the sequence.

(C3) The selection probabilities $p_v(s)$ depend continuously on $v \in \mathcal{Y}$.

(C4) The Metropolis-Hastings proposal distribution for each coordinate i for each kernel P_v is selected from some parametric family whose density function depends continuously on $v \in \mathcal{Y}$.

(C5) The target distribution $\pi(\cdot)$ has continuous density on $\mathcal{X}$.

(C6) The adaptive parameter vector V_{n+1} depends continuously on (some or all of) the chain history $U_0, \ldots, U_n, V_0, \ldots, V_n$.

(C7) There is a deterministic sequence $b_n \searrow 0$ such that the components $V_{n,i}$ of the adaptive parameter vectors $V_n \in \mathcal{Y}$ all satisfy the bound $|V_{n+1,i} - V_{n,i}| \le b_n$.

These conditions all hold for all of the adaptive algorithms used in this paper. Indeed, (C0) holds for all irreducible Metropolis-Hastings kernels (Tierney, 1994, Corollary 2), which includes all the fixed-v kernels considered here since by (10) and (11) each sweep always has a positive probability of including each variable; (C1) holds since the Markov chain and adaption variables are all explicitly defined (see (8) and (12) and (11)) to be uniformly bounded away from 0 and from infinity, so they remain within fixed closed intervals on which condition (C0) continues to hold; (C2) holds by explicit construction of the algorithms, with the selection probabilities $p_v(s)$, indicated as $w_k(b)$ in our algorithm, defined by the adaptive scanning scheme described in Section 4.3.1; (C3) holds since the selection probabilities $w_k(b)$ are defined via (10) in terms of the $\underline{\gamma}_k$ vectors, and furthermore any function on a discrete set like $\{0, 1\}$ is continuous by definition; (C4) holds since the proposal densities used (lognormal, logit, etc.) are all continuous functions of their parameters; (C5) holds since the joint density (3) and likelihood function (4) are continuous functions of their arguments (and, again, any function on a discrete set is continuous by definition); (C6) holds since the adaptive parameters like $s_k(m)$ are continuous functions of the corresponding batch values; and (C7) holds for each coordinate, either explicitly since the amount by which the adaptive parameter is changed goes to 0 as in (9) for the fixed-size batches m, or else because it is defined in terms of empirical means and variances of increasing overlapping batches as is ensured, for example, by the $\sqrt{bS}$-discard defined at the beginning of Subsection 4.3.1 so the differences of means etc. must therefore converge to zero, and furthermore by compactness this convergence must be uniform over all adaptive parameters in $\mathcal{Y}$, as required.

Theorem 1 *Assuming (C0)–(C7), the adaptive algorithm is ergodic, i.e.*

$$\lim_{n\to\infty} \sup_{B\subseteq\mathcal{X}} \Big|\mathbf{P}(U_n \in B \,|\, U_0 = u, V_0 = v) - \pi(B)\Big| \;=\; 0\,, \qquad u \in \mathcal{X},\ v \in \mathcal{Y}\,,$$

and also satisfies a weak law of large numbers (WLLN) for all bounded functionals:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(U_i) \;=\; \pi(h)\,, \qquad h : \mathcal{X} \to [-M, M], \quad \text{some } M < \infty\,.$$

Proof. According to Theorems 5 and 23 of Roberts and Rosenthal (2007), the theorem will follow if we can establish **(a)** the *Simultaneous Uniform Ergodicity* property that for all $\epsilon > 0$, there is $N = N(\epsilon) \in \mathbf{N}$ such that $\|P_v^N(u,\cdot) - \pi(\cdot)\| \le \epsilon$ for all $u \in \mathcal{X}$ and $v \in \mathcal{Y}$; and **(b)** the *Diminishing Adaptation* property that $\lim_{n\to\infty} \sup_{u\in\mathcal{X}} \|P_{V_{n+1}}(u,\cdot) - P_{V_n}(u,\cdot)\| = 0$ in probability. Furthermore, their Corollary 8 states that under (C0) and (C1), property (a) follows if the mapping $(u,v) \mapsto T(u,v,n)$ is continuous for each fixed $n \in \mathbf{N}$, where we let

$$A^{(n)}((u,v), B) = \mathbf{P}[U_n \in B \,|\, U_0 = u, V_0 = v]\,, \qquad B \subseteq \mathcal{X}$$

record the distribution of U_n for the adaptive algorithm, and let

$$T(u,v,n) \;=\; \|A^{(n)}((u,v),\cdot) - \pi(\cdot)\| \;\equiv\; \sup_{B\subseteq\mathcal{X}} |A^{(n)}((u,v),B) - \pi(B)|$$

denote the total variation distance to the target distribution $\pi(\cdot)$.

We first establish property (b). In light of the algorithm's structure (C2), the continuity properties (C3)–(C6) imply continuity of the mappings $v \mapsto P_v(u,\cdot)$ for each fixed $u \in \mathcal{X}$ in the total variation topology (since total variation distance depends continuously on densities). The compactness condition (C1) then implies that this continuity is uniform in u. Hence, property (b) follows from the decreasing differences of the adaptive parameters as in condition (C7).

Next, we decompose the distribution $A^{(n)}((u,v),\,\cdot)$ as

$$A^{(n)}((u,v),\,\cdot) \;=\; r(u,v,n)\, A_s^{(n)}((u,v),\,\cdot) + (1 - r(u,v,n))\, A_m^{(n)}((u,v),\,\cdot)\,,$$

where $r(u,v,n)$ is the probability that at least one of the continuous components of the chain has not yet moved by time n, with A_s the corresponding conditional distribution, and A_m is the conditional distribution of the complementary event. Now, if the chain does not move in one of its continuous components, then it is singular with respect to $\pi(\cdot)$, so

$$T(u,v,n) \;=\; r(u,v,n) \;+\; (1 - r(u,v,n))\, \|A_m^{(n)}((u,v),\,\cdot) - \pi(\cdot)\|\,. \tag{13}$$

To continue, consider two different copies of the adaptive chain, $\{U_n, V_n\}$ and $\{U_n', V_n'\}$. Suppose their initial values satisfy $\|U_0' - U_0\| + \|V_0' - V_0\| < \epsilon$ for some small $\epsilon > 0$. We claim that for each fixed $n \in \mathbf{N}$, there is $d_n(\epsilon)$ with $\lim_{\epsilon\searrow 0} d_n(\epsilon) = 0$, such that the two copies can be coupled in such a way that with probability $\ge 1 - d_n(\epsilon)$, for each coordinate i, either the two copies are identical ($U_{n,i}' = U_{n,i}$), or both copies are still equal to their respective starting values ($U_{n,i} = U_{0,i}$ and $U_{n,i}' = U_{0,i}'$). Indeed, the continuity conditions (C3)–(C6), which each imply uniform continuity by (C1), together imply that the two chains can be coupled so that at each iteration n, with probability which converges to 1 as $\epsilon \searrow 0$, the two chains will each select the same sequence $s \in \mathcal{S}$, the same proposal states in $\mathcal{X}$, the same decisions to accept/reject the proposal states, and the same updated adaptive parameter in $\mathcal{Y}$. The claim follows.

The coupling inequality then implies that $|r(u', v', n) - r(u, v, n)| \leq d_n(\epsilon)$, and also $\|A_m^{(n)}((u, v), \cdot) - A_m^{(n)}((u', v'), \cdot)\| \leq d_n(\epsilon)$. Hence, by (13), $|T(u', v', n) - T(u, v, n)| \leq 3\, d_n(\epsilon)$. This proves the continuity of the mapping $(u, v) \mapsto T(u, v, n)$, and thus establishes property (a), and thus completes the proof of the theorem. □

6. RESULTS

6.1. *Simulation study*

In this section we report the results of the simulation study we perform in order to evaluate the performance of the HESS algorithm, imposing different structures on Ω. We compare our method with three recently proposed algorithms, namely MOM (Kendziorski *et al.* 2006), BAYES (J & X, 2007) and Stochastic Partitioning Algorithm (SPA) (M & T, 2009), discussed in Section 2.2.

To build realistic examples, all six simulated data sets are based on a design matrix X derived from phased genotype data spanning 500-kb, region ENm014, Yoruba population (HapMap project): the data set originally contained 1,218 SNPs (Single Nucleotide Polymorphism), but after eliminating redundant variables, the set of SNPs is reduced to $p = 498$, with $n = 120$, giving a 120×498 design matrix. The benefit of using real data for the X matrix is that the pattern of pairwise correlation, linkage disequilibrium (LD), is complex and hard to mimic and blocks of LD are not artificial, but they derive naturally from genetic forces, with a slow decay of the level of pairwise correlation between SNPs. In all examples, we placed up to six 'hot spots' at SNPs 30, 161, 225, 239, 362 and 466 inside blocks of correlated variables. The first four SNPs are weakly dependent ($r^2 < 0.1$), while the remaining two SNPs are correlated with each other and also linked to SNP 239 ($r^2 \simeq 0.5$), creating potentially a masking effect difficult to detect. The six simulated examples can be summarised as follow:

Sim1: we simulated $q = 100$ responses (transcripts), with the eQTLs at SNP 30 and 239 influencing transcripts 1-20 and 71-80, SNP 161 influencing transcripts 17-20, SNP 225 influencing transcripts 91-100, and finally eQTLs 362 and 466 influencing transcripts 81-90. The goal of this example is to let some transcripts be predicted by multiple correlated markers: for instance transcripts 17-20 are regulated by SNPs 30, 161, 239 at the same time. Altogether 50 transcripts are under genetic control and for these, the effects and the error term are simulated as in J & X (2007) with $\beta_{kj} \sim N\left(0, 0.3^2\right)$ and $\underline{\epsilon}_k \sim N_n\left(\underline{0}, \sigma_k^2 I_n\right)$ with $\sigma_k = 0.1$. All other responses are simulated from the noise.

Sim2: As in the previous example, we simulated 100 responses, but there are only three hot spots (30, 161, 239). Transcripts 81-90 and 91-100 are obtained by a linear transformation of transcripts 20 and 80 using a mild negative correlation (in the interval [-0.5,-0.4]) and a strong positive correlation (in the interval [0.8,0.9]) respectively. The goal of this example is to simulate correlation among some transcripts that is not due to SNPs, creating possible false positive associations.

Sim3: This simulation set-up is identical to the first example for the first 100 responses, but we increase the number of simulated responses to $q = 1,000$, with all additional 900 responses simulated from the noise.

Sim4: As in the second simulated data set for the first 100 responses, with additional 900 responses simulated from the noise, and altogether $q = 1,000$.

Sim5: In this example we simulated $q = 100$ responses with the SNPs-transcripts association similar to the ones described in M & T (2009). We partitioned the 100 transcripts into 10 groups with four of them linked to some combinations of the six

hot spots (30, 161, 225, 239, 362 and 466). Finally the *same* effect is simulated for each of the four partitions from a uniform distribution in $[-5,-2]\bigcup[2,5]$ with $\underline{\epsilon}_k \sim N_n\left(\underline{0},\sigma_k^2 I_n\right)$ with $\sigma_k = 1$.

Sim6: The same groups as in **Sim5** are used in this example, but, irrespectively of the SNPs-transcripts partition structure, the effects and the error terms are simulated as in J & X (2007) with $\beta_{kj} \sim N\left(0, 0.3^2\right)$ and $\underline{\epsilon}_k \sim N_n\left(\underline{0},\sigma_k^2 I_n\right)$ with $\sigma_k = 0.1$. In this final example, the unrealistic assumption of 'blocks of similar effects' is removed and the signal to noise ratio is lower than the one implemented in M & T (2009).

Sim1 and **Sim2** will be used to compare HESS and BAYES; **Sim5** and **Sim6** to compare HESS to SPA. On **Sim3** and **Sim4**, we will compare HESS to MOM and explore adaptive scanning strategies.

6.2. *Postprocessing*

To illustrate the performance of HESS, we report results with a burn-in of 1,000 sweeps and a run length of 2,000 sweeps. m batches are of length 50 and we increment the b batch index every 50 sweeps. Adaptation for the Ω updates starts at the beginning, while if the adaptive scanning version is implemented, adaptation of the $w_k(b)$ starts at the end of the burn-in. We run 3 chains ($L = 3$) and stop temperature adaptation at the end of the burn-in. We set the hyper-parameters $a_{\omega_{kl}}$ and $b_{\omega_{kl}}$ so that $E(p_{\gamma_{kl}}) = V(p_{\gamma_{kl}}) = 2$, $\forall k, l$ if model (iii) for Ω is chosen and $c_{\omega_{jl}} = d_{\omega_{jl}} = 0.05$ $\forall j, l$ if model (ii) is preferred. All the results presented for **Sim1**-**Sim2** and **Sim5**-**Sim6** were run with the fixed fraction ϕ scanning, $\phi = 0.25$.

Amongst the rich posterior output produced by HESS, we will focus on ρ_j (model (iii)) or ω_j (model (ii)) in order to characterise hot spots. We will also present summaries of γ_{kj} for MAP configurations. It is not our purpose here to discuss in depth a variety of classification rules that can be built to 'declare' a predictor as a hot spot, as this would require a separate study. For model (iii), in the spirit of cluster detection rules in disease mapping (Richardson *et al.*, 2004), we will use tail posterior probabilities of the propensities ρ_j, i.e. declare the j^{th} predictor to be a hot spot if $Pr(\rho_j > 1\,|Y\,) > 0.8$. We use a 2-components mixture of beta distributions to analyse the posterior distribution of the column effects: ω_j in model (ii) and ζ_j in J & X. This mixture has typically a component with a high peak around small values that can be interpreted as representing the background rates. We will declare the j^{th} predictor to be a hot spot if the associated weight of the background component is small, say less than 0.2. Thresholds can be determined for specified FDR if so required.

6.3. *Comparison of HESS and BAYES (Jia and Xu) on **Sim1** and **Sim2***

On Figure 1, we present a summary output of the run of BAYES $E(\zeta_j\,|\,Y)$ (left), model (ii) $E(\omega_j\,|\,Y)$ (middle), and model (iii) $E(\rho_j\,|\,Y)$ (right) on the on **Sim1** and **Sim2** set-ups. Results of 5 replications are represented. We first remark that BAYES is not performing well on **Sim1**, in particular marker 225, and 362 and 466 (those with potential masking) are not detected as hot spots in 2, 3 and 2 (resp) of the 5 replicates. We also see that there is some difficulty in separating the background rates of ζ_j from those of the true hot spots and that there are a number of false positives hot spots being detected (crosses) particularly in **Sim2** around marker 30 and 161. When investigating in more detail the runs of BAYES, we found evidence that the BAYES algorithm does not always mix adequately and that it can get stuck in local modes, creating false associations. For example, the

SNP-response associations responsible for the false positive hot spots near SNP 161 were incorporated by BAYES algorithm early on and remained throughout during the MCMC run. This is not unexpected since BAYES uses only Gibbs sampling to perform variable selection without integrating the regression coefficients and that single variable updates can lead to poor mixing when the predictors are correlated. In contrast, both HESS models (ii) and (iii) find all the hot spots in **Sim1** and only miss one in one replicate of **Sim2**. The benefits of the multiplicative model (iii) in terms of clear separation of the hot spots from the background are clearly visible. The additional sparsity of model (iii) has led to a useful shrinkage of the background rate, and values of $E(\rho_j \mid Y)$ give a decisive indication of high propensity for the true hot spots markers. On the other hand, for model (ii) we observe more variability of the background rate leading to difficulties of classification and potentially more false positives, in line with our intuition that model (ii) might over estimate hot spot probabilities. In view of this and other experiments that we have carried out, we will focus our reporting on model (iii).

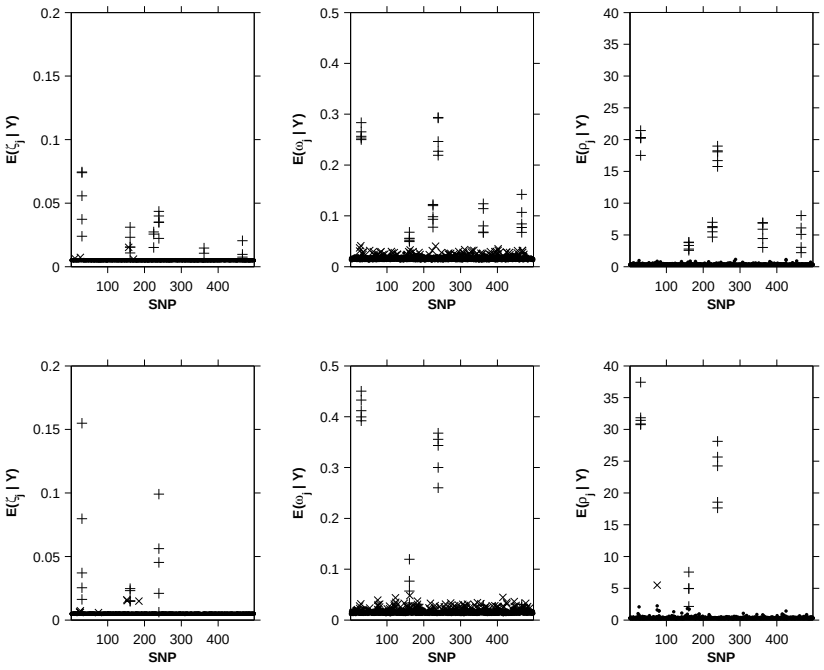

Figure 1: *Detection of hot spots: comparison of the performance of BAYES (left) and HESS: model (ii) (middle), model (iii) (right).* + *true positive,* × *false positive, all other values are indicated with a black dot.*

6.4. *Comparison of HESS and SPA (M & T) on **Sim5** and **Sim6***

On Figure 2, we present an example from one of the replicates of **Sim5** and **Sim6** highlighting the general pattern of results. The blocks of simulated effects are represented on the left, the MAP output given by Stochastic Partitioning Algorithm (SPA) in the middle and a summary of the posterior frequency: $E(\gamma_{kj} \,|\, Y)$ for the MAP configuration produced by HESS on the right. Recall that **Sim5** was simulated with common effects within blocks, following the simulation set-up of M&T. Nonetheless, some blocks of effects are not detected by SPA (e.g. at SNP 30, 239 and 466 in the upper part of the plot) and there is evidence also of a false positive block at SNPs 75 and 159. The **Sim6** setting with effects varying within blocks and lower signal to noise ratio induces the SPA to split into many atomic subsets to accommodate the variability of effects, no information can be borrowed and some effects are not detected. When running SPA, we found the tuning of their partition parameter ρ quite difficult, with results highly sensitive to changes in ρ. As recommended by M & T, we attempted to balance the two types of reversible jump moves by trying different values of ρ and the results reported achieved a balance of nearly 50% over 10^6 iterations.

The results of HESS are consistent with different signal to noise ratio between **Sim5** and **Sim6**. In **Sim5** all blocks of effects are detected with high probability. In **Sim6**, some weaker effects are missed, but altogether, the general pattern of the blocks is clearly apparent, and there are few false positives. Hence, the multiplicative model (iii) gives not only a good tool for detecting hot spots as shown in 6.3, but also a rich output that can be used to finely discover pairwise associations between responses and predictors, irrespective of an imposed block structure on the effects.

6.5. *Comparison of HESS and MOM on **Sim3** and **Sim4***

In these two set-ups, the number of responses is substantially increased to $q = 1,000$, with only 50 responses truly associated to the markers. In Figure 3 first column, we report the results of 5 replicates, focussing on the comparison of the posterior probabilities of hot spots that can be obtained by running respectively the MOM algorithm and HESS model (iii). We first point out that our simulation set-up is quite different to that of Kendziorski *et al.* (2006) in that (i) we have a smaller absolute number of transcripts associated with the markers (50 in our case and between 500 and 1,500 in their case), even though the fraction of responses associated are comparable (5% versus 3%), and (ii) we are considering about 500 predictors instead of 23. Hence the mixture identification underlying MOM has less information, and could be expected to have more instability. We observe that overall both methods find easily the great majority of the hot spots (indicated by +) as the respective posterior probabilities are located in the top right corner. The notable difference is the clear separation of $Pr(\rho_j > 1 \,|\, Y)$ between associated and non-associated markers, with a clump of low values (below < 0.4) for most non-associated markers, whereas the posterior probabilities for hot spot provided by MOM are more spread out, with some values close to 1 for non-associated markers in particular in **Sim4**.

6.6. *Adaptive scanning*

We also use the set-ups of **Sim3** and **Sim4** to investigate the performance of our adaptive scanning algorithm. One important tuning parameter in our adaptation scheme is the constant c in (11) that controls how fast the $\varepsilon(b)$ will adapt. We

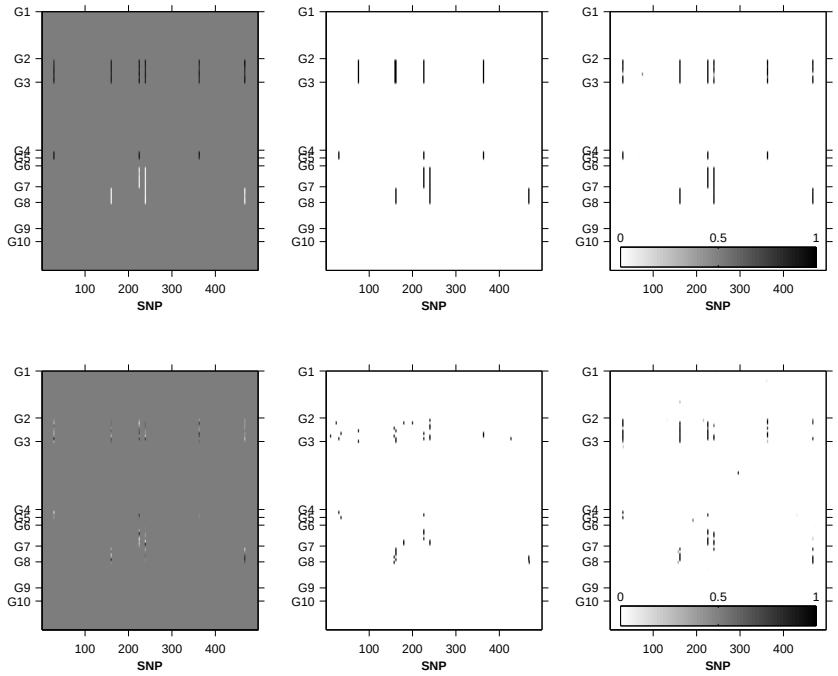

Figure 2: *Comparison of HESS model (iii) with SPA model. True simulated effects(left), SPA MAP configuration (middle), Posterior frequencies corresponding to the MAP configuration of HESS (right).*

explored several choices: $c = 10, 10^2, 10^3$, and in our limited experiments, found that $c = 100$ provided a good compromise. Figure 3 (middle column) displays the selection probabilities, $w_k(b)$ for one adaptive run of **Sim3** and **Sim4**, where adaptation starts after 1,000 burn-in sweeps (i.e. at batch index $b_0 = 20$), with $S = 50$. It is clear that for the 50 associated responses (in black), $w_k(b)$ grows nicely reaching a ratio of 3 to 1 after 60 batch updates (3,000 sweeps). On the other hand, the majority of non-associated responses (light grey) have a decreasing $w_k(b)$. It is further interesting to see that 'recovery' is happening. For example in the bottom plot, one of the associated response started with decreasing $w_k(b)$, but at batch 30, this trend was reversed. Similarly, some of the non-associated responses that have increasing $w_k(b)$ at the start show turning points where this trend is reversed, indicating that the chosen adaptive scheme has viable elasticity in a short number of batch updates. The right column of Figure 3 compares tail posterior probabilities of hot spots between adaptive and non-adaptive scanning version and shows that there is excellent agreement for the hot spot probabilities (shown with +); hence adaptive and non-adaptive scanning schemes converge to similar posteriors as should be expected from the theory. A small improvement regarding the dispersion of the tail probabilities of the adaptive scanning scheme for

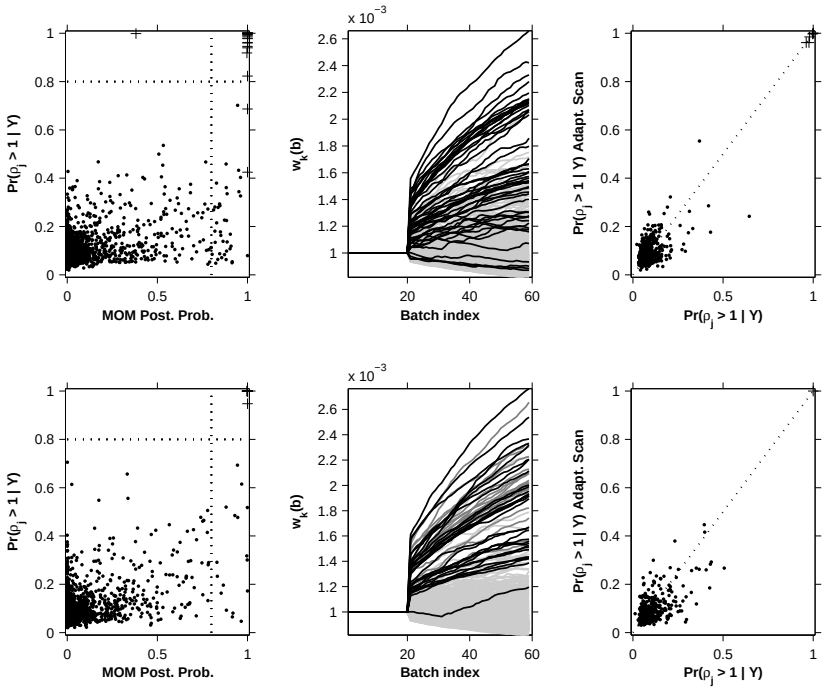

Figure 3: *Left column: Comparison of HESS and MOM. Middle column: Selection probabilities $w_k(b)$ for HESS with adaptive scanning (black: associated responses, light grey: non-associated responses). Right column: Comparison of tail posterior probabilities for adaptive and non-adaptive versions. Output from one simulation of* **Sim3** *(top) and* **Sim4** *(bottom).*

non-associated responses is also suggested. Note that starting the adaptive scheme at the end of a burn-in of 1,000 sweeps is quite conservative as, by then, the two algorithms have already homed in on the interesting parts of the model space.

To illustrate more clearly the benefits of adaptive scanning, we carried a further experiment, starting the adaptive scanning after only 100 sweeps (i.e. $b_0 = 2$) on **Sim3**. To make comparison easier between adaptive and non-adaptive scanning, we fix the value of g in both algorithms to the unit information prior, i.e. $g = n$. Figure 4 (top left) shows again how the $w_k(b)$ start increasing, almost immediately for most of the associated responses. This time, there is more difference in the tail posterior probabilities, with higher values overall for the associated responses (+), and less dispersion for the non-associated responses (Figure 4, top middle for the adaptive scanning). In complement, we monitored the fraction of misclassified γ_{kj} as the two algorithms progress (Figure 4, top right). We see that the adaptive scanning has a steeper rate of misclassification decrease than the non-adaptive version, indicating that it learns faster the correct associations.

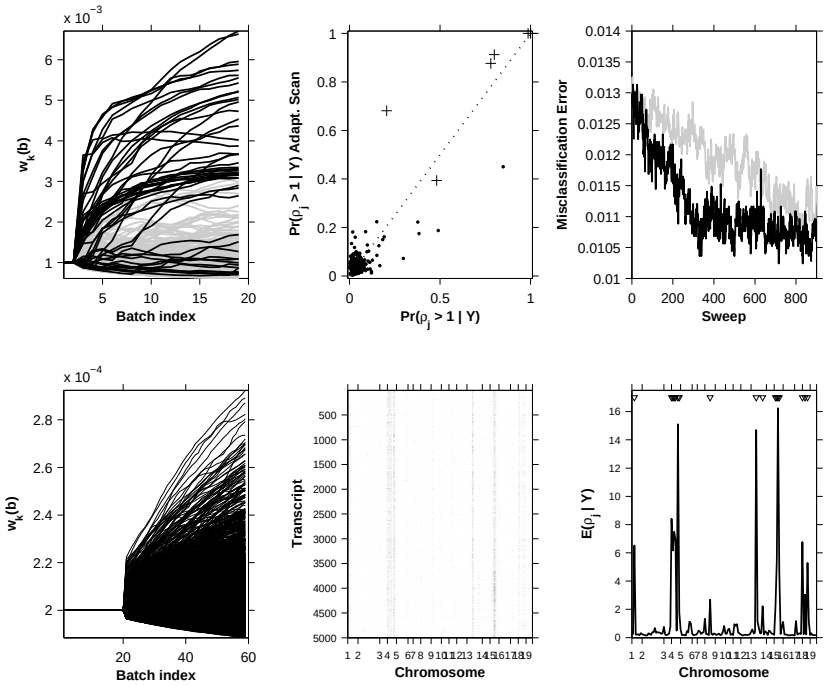

Figure 4: *Top: Comparison of adaptive versus non-adaptive scanning HESS algorithms on* **Sim3** *when adaptive scanning starts after 100 sweeps. Left: Selection probabilities* $w_k(b)$ *(black: associated responses, light grey: non-associated responses). Middle: Tail posterior probabilities. Right: Misclassification error (black: adaptive, light grey non-adaptive). Bottom: eQTL analysis of* F_2 *mice. Left: Selection probabilities,* $w_k(b)$*. Middle: Posterior frequencies of* γ_{kj}*. Right: Posterior propensity of hot spot (SNPs with associated tail posterior probabilities above 0.8 are indicated with triangles.)*

eQTL analysis of data from a study of diabetes in F_2 *mice.* Finally, we performed an e-QTL analysis on publicly available data arising from an experiment investigating genetic causes of obesity and diabetes, data that were previously analysed by Kendziorksi *et al.* (2006) and Jia and Xu (2007). The data set comprise 60 F_2 *ob/ob* mice segregating for phenotypes associated with diabetes and obesity, on which $p = 145$ markers were recorded. Gene expression was measured by Affymetrix Gene Chips (MOE43A,B), and for this illustrative example, we analyse the top $q = 5,000$ most varying transcripts. The adaptive scanning HESS (with fixed $g = 60$) was used to analyse this data and the Matlab code run on a 3GHz CPU with 4Gb RAM desktop took 67 hours to complete. The bottom part of Figure 4 shows some of the posterior output. On this challenging joint analysis, we see again that some of the selection probabilities have a marked increase. The posterior expectation of ρ_j give clear indication of several hot spots and the posterior frequencies of γ_{kj} characterise

further the associated responses. Using the tail probability rule, we would declare 17 hot spots on this data set. In particular, there are three massive hot spots in chromosome 4, SNP D4Mit186, chromosome 13, SNP D13Mit91, and chromosome 15, SNP D15Mit63.

7. DISCUSSION

We have presented new models and algorithms for regression analysis of a large number of responses and a large number of predictors. We have shown that by comparison to currently proposed models and algorithms, our implementation performs better in a variety of situations. We found that the new multiplicative model for the joint probability allows an excellent separation between hot spot and background, and we would recommend to use this formulation rather than the simple column effect model. Hierarchical extensions of the multiplicative model could be considered, which would treat the $(\{\rho_{jl}\}, 1 \leq j \leq p, 1 \leq l \leq L)$ as random effects, coming, say, from an exchangeable or a mixture prior. These extensions are certainly worth considering, but as p is large, will require to develop new efficient updating strategies for the set of ρ_j.

Stimulated by the goal to make fully Bayesian joint analysis more computationally feasible, in this paper, we provide an important proof of concept for a class of adaptive scanning strategies and discuss in details one implementation of such a scheme. Theoretical conditions for ensuring convergence are derived that are relatively easy to satisfy and leave many degrees of freedom to the MCMC designer. The key ingredients are the definition of the batch with the need for increasing overlap and the formulation of the quantities on which to base the adaptation. The amount of information needed to be accumulated before the start of the adaptation is also an important feature where gains of efficiency could be expected, in line with one of our experiments. We stress that the results that we show only cover a small aspect of the potential improvements that will be derived from such schemes and that extensive experimentation is now required in order to give guidelines on these choices. In conclusion, we believe that adaptive strategies, in particular adaptive scanning, will be very useful in bringing fully Bayesian analyses to integrative genomics in the near future.

REFERENCES

Andrieu, C. and Moulines, E. (2006). On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. *Ann. Appl. Prob.* **16**, 1462–1505.

Bottolo, L. and Richardson, S. (2010). Evolutionary Stochastic Search for Bayesian model exploration. To appear in *Bayesian Analysis*.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Science* **19(1)**, 81–94.

George, E. I. and McCullogh, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **85**, 398–409.

Jia, Z. and Xu, S. (2007). Mapping Quantitative Trait Loci for expression abundance. *Genetics* **176**, 611–623.

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical methods for expression Quantitative Trait Loci (eQTL) mapping. *Biometrics* **62**, 19–27.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā B* **60**, 65–81.

Monni, S. and Tadesse, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis* **4(3)**, 413–436.

O'Hara, R. B., and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* **4(1)**, 85–118.

Ntzoufras, I. (1999). Aspects of Bayesian variable selection using MCMC. Unpublished PhD Thesis, Athens University.

Richardson, S., Thomson, A., Best, N. and Elliott, P. (2004). Interpreting posterior relative risk estimates in disease mapping studies. *Env. Health Persp.*, **112**, 1016–25.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive MCMC. *J. Appl. Probability* **44**, 458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *J. Comp. Graphical Statist.* **18(2)**, 349–367.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701–1762.

Yi, N. and Shriner, D. (2008). Advances in Bayesian multiple Quantitative Trait Loci mapping in experimental crosses. *Heredity*, **100**, 240–252.