Efficient inference for genetic association studies with multiple outcomes

Hélène Ruffieux[a,b,*], Anthony C. Davison[b], Jörg Hager[a], Irina Irincheeva[a]

[a]Nestlé Institute of Health Sciences SA, Lausanne, Switzerland

[b]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

March 22, 2017

Abstract

Combined inference for heterogeneous high-dimensional data is critical in modern biology, where clinical and various kinds of molecular data may be available from a single study. Classical genetic association studies regress a single clinical outcome on many genetic variants one by one, but there is an increasing demand for joint analysis of many molecular outcomes and genetic variants in order to unravel functional interactions. Unfortunately, most existing approaches to joint modelling are either too simplistic to be powerful or are impracticable for computational reasons. Inspired by Richardson et al. (2010, Bayesian Statistics 9), we consider a sparse multivariate regression model that allows simultaneous selection of predictors and associated responses. As Markov chain Monte Carlo (MCMC) inference on such models can be prohibitively slow when the number of genetic variants exceeds a few thousand, we propose a variational inference approach which produces posterior information very close to that of MCMC inference, at a much reduced computational cost. Extensive numerical experiments show that our approach outperforms popular variable selection methods and tailored Bayesian procedures, dealing within hours with problems involving hundreds of thousands of genetic variants and tens to hundreds of clinical or molecular outcomes.

Key words: High-dimensional data; Molecular quantitative trait locus analysis; Sparse multivariate regression; Statistical genetics; Variable selection; Variational inference.

1 Introduction

Much current research in genetics focuses on combining heterogeneous data for the same samples. This is prompted by the increasing availability of diverse molecular data types from a single study and should lead to a more complete understanding of biological systems, as combined inference for data from different molecular layers might unravel regulatory interactions within and across layers (Civelek and Lusis, 2014). An example is *protein quantitative trait locus* (pQTL) analyses to detect associations between hundreds of thousands of genetic variants and hundreds of proteomic expression levels. When associated with disease genetic variants, proteins are often regarded as intermediate phenotypes or molecular proxies for the disease of interest, as they may provide direct

---

*To whom correspondence should be addressed (helene.ruffieux@rd.nestle.com).

insights on biological processes underlying the clinical condition, and likewise for other molecules such as genes or metabolites involved in so-called eQTL or mQTL analyses.

An important goal of molecular QTL studies is to detect genetic variants with *pleiotropic effects*, i.e., variants that regulate the expression levels of several molecules, as the genome location where they lie may initiate essential functional mechanisms (Breitling et al., 2008). As pleiotropy has been acknowledged as a central property of genetic variants causing phenotypic variation, efforts are being made to uncover its activity patterns and gain understanding of the shared biological processes it induces (Sivakumaran et al., 2011; Solovieff et al., 2013). It is also of interest to identify those cases where the same molecule is simultaneously influenced by several genetic variants. This dual task requires a model that allows flexible selection, ideally performed jointly on the genetic variants and the molecular outcomes.

Although in practice univariate analyses still dominate, several proposals for joint modelling of multiple outcomes in genetic association problems have recently been made. Flutre et al. (2013) and Zhou and Stephens (2014) model the outcomes as multivariate responses having a matrix-variate distribution. The former rely on a Bayes factor framework to uncover the associations between a given genetic variant and any subset of outcomes, whereas the latter propose a linear mixed model whose random effect accounts for relatedness among individuals. While these methods show improvements over the fully marginal regression approach, they are restricted to problems with a few outcomes, because their models involve unstructured covariance matrices. O'Reilly et al. (2012)'s MultiPhen method reverses the classical regression setup and fits a succession of models where each genetic variant is regressed on several outcomes. This eliminates the need to model large covariance matrices, but also entails a marginal treatment of the genetic variants. Moreover, MultiPhen does not penalize model complexity, which may cause instabilities when many outcomes are modelled. Molecular QTL problems are particularly complex, because in addition to the so-called $p \gg n$ paradigm, whereby the number of covariates (genetic variants) p greatly exceeds the number of samples n, there is also a "large d" characteristic, as the number of responses (expression levels) d is usually large. Methodologies to accommodate this are needed, especially when joint response modelling is sought.

Two-stage procedures are natural approaches to association problems with both p and d large. 2HiGWAS (Jiang et al., 2015) is essentially an implementation of the screening strategy of Fan and Lv (2008) in the context of longitudinal outcomes. It consists of a dimension reduction step, where each genetic variant is tested against each outcome, followed by a penalized regression recast into functional mapping in which only the genetic variants remaining after screening are involved. While the second stage of 2HiGWAS is an interesting approach to joint covariate and response modelling, the method is not tailored to typical molecular QTL analysis, as it is designed for outcomes measured over time. Also, the fully marginal first stage screening, if too stringent, may cause important predictors to drop out of the analysis and thus lead to false negatives. Instead of pruning the covariate set, Wang et al. (2016) summarize information at outcome level in the context of eQTL analyses. They precluster the expression levels using a block-mixture model and test for association between each genetic variant and the resulting clustering. This approach requires that the stability of the clustering and its functional relevance are carefully checked, as the group pattern chosen is critical to subsequent analysis. More generally, two-stage procedures often rely on ad-hoc thresholding decisions at the first stage which influence the conclusions of the

second stage.

The approaches sketched above use very diverse strategies to model predictor and outcome variables in high-dimensional settings. Trade-offs between realistic modelling and computational efficiency are inevitable, but two components seem critical to practical and powerful analyses: involving all variables in a single multivariate model, and maintaining interpretable inference. Unified selection of genetic variants and associated outcomes is then possible, unlike for most existing methods, whose focus is on selecting either predictors or outcomes. The Bayesian framework seems particularly suitable, as it offers flexible modelling possibilities in which biological beliefs may be naturally incorporated. The hierarchical regression approach of Richardson et al. (2010), HESS (hierarchical evolutionary stochastic search), is an appealing example of such approaches; it can identify associations between hundreds of covariates and up to a few thousand responses from a single model. Jia and Xu (2007) and Scott-Boyer et al. (2012) propose methods called BAYES and iBMQ (integrated Bayesian hierarchical model for eQTL mapping) based on models similar to that of HESS, but a major drawback of all three approaches is the lack of scalability of their Markov chain Monte Carlo inference procedures. Problems whose size corresponds to actual genome-wide association studies with molecular outcomes (with hundreds of thousands of genetic variants and hundreds to thousands of outcomes and individuals) are out of reach even for HESS, which is based on adaptive parallel tempering/evolutionary Monte Carlo techniques. We are unaware of any fully multivariate approach that can deal with such data within a reasonable time.

In this paper, we propose a Bayesian inference strategy that avoids sampling, via an algorithm that is fast and whose convergence is easy to monitor, while having performance comparable with Markov chain Monte Carlo approaches. We describe a variational inference procedure for a model similar to that of Richardson et al. (2010), comprising a series of parallel linear regressions, one for each response, combined in a hierarchical manner to leverage shared information. A spike-and-slab prior (Ishwaran and Rao, 2005) is used to induce sparsity of the regression coefficients, and the probability that a given covariate affects any response is modelled through a parameter that is shared across responses. Interpretable posterior quantities, such as the probability of association of each covariate-response pair, are produced. An efficient algorithm for this model is crucial, as its number of parameters can be very large. Our variational approach can update the parameters jointly in a tractable manner. Carbonetto and Stephens (2012) provide a good discussion of variational inference in the context of genetic association and propose a variational regression method called "varbvs", which can be seen as a single-response counterpart of our approach. In our multiple response setting, we show by simulation that our procedure is accurate and reliable, and is better than existing methods at selecting variables for very large problems. Hence, it offers clear added-value in practice: it enables complex and flexible Bayesian inference based on large batches of genetic data, without having to prune them beforehand.

The paper is organized as follows. Section 2 describes the model, discusses its relation to earlier proposals, and presents a procedure to allow for sparsity control at both covariate and response levels. Section 3 gives an overview of variational Bayes approaches and describes our inference strategy. Section 4 compares variational and Markov chain Monte Carlo inferences on the same model, also using direct approximations of posterior quantities. Section 5 describes numerical experiments for larger problems, comparing our method with several predictor selection methods, including the varbvs approach of Carbonetto and Stephens (2012), and with methods

performing combined covariate and response selection, namely HESS (Richardson et al., 2010) and iBMQ (Scott-Boyer et al., 2012). The section also presents a permutation-based comparison of our method with varbvs on a real mQTL problem. Section 6 summarizes the discussion and highlights further possible developments.

Although the applicability of our method is not restricted to any particular context, all the numerical experiments presented in this paper use settings tailored to genome-wide association studies. The data-generation schemes are designed to embody common biological assumptions and are described in Section 5.1. The method is implemented in the publicly available R package `locus`.

2   Model and earlier proposals

Let $y = (y_1, \ldots, y_d)$ be a $n \times d$ matrix of d centered responses and let X be a $n \times p$ matrix of p covariates, for each of n samples. The covariates are p genetic variants, more precisely single nucleotide polymorphisms, SNPs, and the responses might represent d gene, protein or metabolite expression levels for n individuals, depending on whether an eQTL, pQTL or mQTL problem is considered. Our model is intended to accommodate all the constraints entailed by molecular QTL analyses; it is adapted from that of HESS (Richardson et al., 2010). Suppose that

$$
\begin{aligned}
y_t &\mid \beta_t, \tau_t \sim \mathcal{N}_n\left(X\beta_t, \tau_t^{-1}I_n\right), & \tau_t &\stackrel{\text{ind}}{\sim} \text{Gamma}(\eta_t, \kappa_t), & t &= 1, \ldots, d, \\
\beta_{st} &\mid \gamma_{st}, \tau_t, \sigma^2 \sim \gamma_{st}\,\mathcal{N}\left(0, \sigma^2\,\tau_t^{-1}\right) + (1 - \gamma_{st})\,\delta_0, & \sigma^{-2} &\sim \text{Gamma}(\lambda, \nu), & \\
\gamma_{st} &\mid \omega_s \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\omega_s), & \omega_s &\stackrel{\text{ind}}{\sim} \text{Beta}(a_s, b_s), & s &= 1, \ldots, p,
\end{aligned}
$$

where δ_0 is the Dirac distribution. Each response, y_t, is related linearly to the covariates and has a specific precision, τ_t. The responses are conditionally independent across the regressions, but dependence among responses associated with the same covariates is captured through the prior specification of parameters ω_s and σ^{-2}, which are shared across the responses. This formulation circumvents modelling the covariance between the responses, which is infeasible when d is large. Each covariate-response pair has its own regression parameter, β_{st}, for which sparsity is induced using a spike-and-slab prior. The binary parameter γ_{st} acts as a "pair selection" indicator; covariate X_s is associated with response y_t if and only if $\gamma_{st} = 1$. The parameter σ represents the typical size of nonzero effects and is modulated by the residual variance, τ_t^{-1}, of the response concerned by the effect. The parameters $\gamma_{s1}, \ldots, \gamma_{sd}$ specify the response(s) associated with X_s and are identically distributed as Bernoulli with common parameter ω_s. Thus, ω_s controls the proportion of responses associated with covariate X_s. The goal of inference is variable selection. Selection of predictors can be performed by ranking the posterior means of $\{\omega_s\}$ and selection of covariate-response pairs can be performed by ranking the posterior probabilities of inclusion, i.e., the posterior means of $\{\gamma_{st}\}$.

Our model differs from that of Richardson et al. (2010) in two respects. One concerns the treatment of the regression coefficient parameters β_{st}: we use independent priors, whereas Richardson et al. rely on g-priors (Zellner, 1986). The main motivation for our choice is that the effects of genetic variants on a given outcome can be understood as causal, since no retroactive process can affect the variants, and they can take place at locations of the genome that are far apart, so their correlation structure need not reflect the spatial correlation of the SNPs; see Guan and Stephens (2011). Jia and Xu (2007) also rely on independent priors for the regression coefficients of BAYES,

but they model the latter with a mixture of two normal distributions rather than a spike-and-slab prior and impose a residual variance parameter that is common to all responses. This stringent assumption may represent a weakness of their proposal.

The second difference concerns the third level of the model. Richardson et al. (2010) opt for a quite complex specification, in which

$$\omega_{st} = \rho_s \omega_t \,, \qquad \omega_t \sim \text{Beta}(a_t, b_t)\,, \qquad \rho_s \sim \text{Gamma}(c_s, d_s)\,, \qquad 0 \leq \omega_{st} \leq 1\,. \quad (1)$$

In their case, the inclusion probability of X_s for response y_t is modelled through ω_t; it is specific to that response but can be regulated using the parameter ρ_s, common to all responses. Jia and Xu (2007) and Scott-Boyer et al. (2012) propose other variants for this prior. The former choose a treatment similar to ours, with $\omega_{st} \equiv \omega_s \sim \text{Dirichlet}(1,1)$, and the latter consider an additional level of hierarchy,

$$\omega_{st} \mid a_s, b_s, \pi_s \sim \pi_s \delta_0 + (1 - \pi_s)\text{Beta}(a_s, b_s)\,, \qquad \pi_s \sim \text{Beta}(a_0, b_0)\,, \quad (2)$$

with $a_s \sim \text{Exp}(\lambda_a)$ and $b_s \sim \text{Exp}(\lambda_b)$. Our choice $\omega_{st} \equiv \omega_s \sim \text{Beta}(a_s, b_s)$ is partly driven by our wish to design a simpler model and partly by practical considerations, since it ensures a closed form for our variational algorithm, unlike with (1). While such a formulation was mentioned by Richardson et al. (2010) and by Scott-Boyer et al. (2012), they did not pursue it because of concerns regarding its ability to control for multiplicity. Indeed, our model inherently enforces sparse associations as the number of responses, d, increases, but no control is achieved when the number of covariates, p, grows. We address this below by providing a procedure to induce a correction through the prior of ω_s.

Part of the flexibility of our model comes from the fact that the hyperparameters, $a, b \in \mathbb{R}_+^p$ (for ω's Beta prior), $\lambda, \nu \in \mathbb{R}_+$ (for σ^{-2}'s Gamma prior) and $\eta, \kappa \in \mathbb{R}_+^d$ (for τ's Gamma prior) are readily interpreted. One option is to set them based on external information regarding the likelihood of given associations, if available. For instance, to favour associations with covariate X_s, one can set a_s and b_s so that the prior proportion of responses affected by X_s, $\text{E}(\omega_s)$, is large. The use of such assumptions may be very efficient, but it may also skew the inference towards existing knowledge. In the simulations presented in this paper, we assume that the regression and variance parameters are exchangeable, i.e., that all covariates, and responses have the same prior propensity to be involved in associations, by selecting a single value for all components of a, b, η and κ. Without favouring any covariate or response, however, we can control signal sparsity at the level of covariates by specifying (possibly through cross-validation) a prior average number of covariates, p^*, expected to be included in the model. Setting

$$a_s \equiv 1\,, \qquad b_s \equiv d(p - p^*)/p^*\,, \qquad 0 < p^* < p\,, \quad (3)$$

the prior probability that X_s is associated with at least one response is

$$p\left(\cup_{t=1}^d \{\gamma_{st} = 1\}\right) = 1 - \frac{\prod_{j=1}^d (b_s + d - j)}{\prod_{j=1}^d (a_s + b_s + d - j)} = \frac{p^*}{p}\,,$$

and simpler models are favoured as p increases. To see this, one can consider the prior odds ratio representing the support for a model to have an additional response associated with X_s, i.e.,

$$\text{POR}(q_s - 1 : q_s) = \frac{p\left(\sum_{t=1}^d \gamma_{st} = q_s - 1\right)}{p\left(\sum_{t=1}^d \gamma_{st} = q_s\right)} = \frac{b_s + d - q_s}{a_s + q_s - 1}\,, \qquad q_s = 1, \ldots, d\,. \quad (4)$$
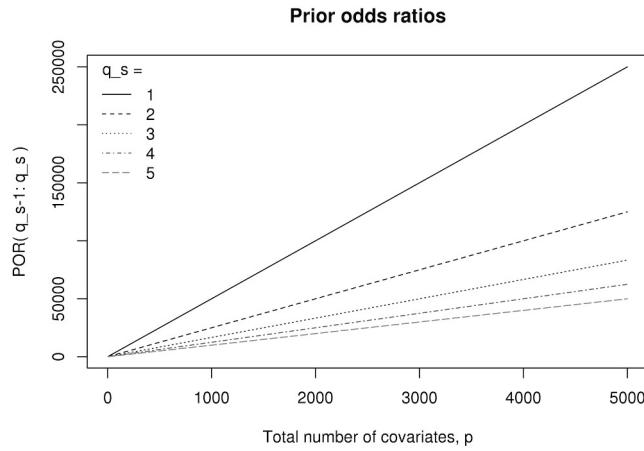
Figure 1: Prior odds ratios, $\text{POR}(q_s - 1 : q_s)$, for $q_s = 1, \ldots, 5$, a_s and b_s as in (3), $d = 100$, $p^* = 2$, and for a total number of covariates ranging from $p = 5$ to $5,000$; see Scott and Berger (2010) for a similar visualization of prior odds ratios in a single response context.

Table 1: Multiplicity adjustment at covariate level. The mean numbers of false positives (FP) and true positives (TP) obtained with the uncorrected and corrected regimes are compared for $p_0 = 20$ active covariates and an increasing number of noise covariates, $p - p_0$. The total number of responses is $d = 25$. 64 replicates were performed; standard errors are in parentheses.

p	50	250	500	1,000	2,500
Mean # of FP					
Uncorrected	0.61 (0.66)	5.23 (2.24)	10.58 (3.16)	22.38 (4.48)	52.61 (8.39)
Corrected	0.77 (0.81)	0.70 (0.99)	0.61 (0.77)	0.39 (0.58)	0.44 (0.59)
Mean # of TP					
Uncorrected	19.95 (0.21)	20.00 (0.00)	19.98 (0.12)	20.00 (0.00)	19.94 (0.24)
Corrected	19.97 (0.18)	19.91 (0.34)	19.81 (0.43)	19.77 (0.56)	19.38 (0.86)

Clearly, penalties arise and increase with the total number of responses in the model, d. Figure 1 displays (4) for $q_s = 1, \ldots, 5$ as a function of p and indicates that, when a_s and b_s are specified as in (3), the penalties also increase with the total number of covariates, p, therefore naturally adjusting for multiplicity. Moreover, the penalties are not uniform when moving from one to two responses associated with X_s, or from four to five, for instance.

The experiment reported in Table 1 confirms that adjustment takes place in practice. It considers problems with $p_0 = 20$ "active" covariates, i.e., those associated with at least one response, and an increasing number of "noise" covariates and it compares the regime with a_s and b_s set according to (3) to an "uncorrected" regime with $a_s \equiv 1$, $b_s \equiv 2d - 1$, so that the prior mean number of responses associated with X_s is 0.5, i.e., $\text{E}(\omega_s) \equiv (2d)^{-1}$. The number of false positives, based on a posterior probability of inclusion greater than 0.5, grows linearly with p when the uncorrected model is used but remains roughly constant close to zero with correction (3), giving a clear multiplicity adjustment. Other experiments confirmed strong sparsity control; the reported findings on real data should therefore be plausible when (3) is used.

3 Variational inference

Section 2 described some differences between our model and those of Jia and Xu (2007, BAYES), Richardson et al. (2010, HESS) and Scott and Berger (2010, iBMQ), but a more fundamental distinction concerns the inference procedure. The three earlier methods rely on Markov chain Monte Carlo (MCMC) techniques and require massive computing resources when the dimensionality of the problem is large. We instead employ a variational inference procedure, which is deterministic and hence can be much cheaper (Ormerod and Wand, 2010).

Instead of sampling from the joint posterior probability $p(\theta \mid y)$ of the parameter vector of interest, θ, variational approaches proceed by replacing it by a tractable analytical approximation, $q(\theta)$. We focus on so-called *mean-field* variational formulations (Xing et al., 2002; Attias, 2000) to construct such a class, i.e., we assume that $q(\theta)$ factorizes over some partition of θ, $\{\theta_j\}_{j=1,\dots,J}$,

$$q(\theta) = \prod_{j=1}^{J} q_j(\theta_j);$$

no further assumption is made about the distribution, and in particular no constraint is imposed on the functional forms of the $q_j(\theta_j)$. Here, we consider the factorization

$$q\left(\beta, \gamma, \tau, \sigma^{-2}, \omega\right) = \left\{\prod_{s=1}^{p}\prod_{t=1}^{d} q(\beta_{st}, \gamma_{st})\right\} \left\{\prod_{s=1}^{p} q(\omega_s)\right\} \left\{\prod_{t=1}^{d} q(\tau_t)\right\} q\left(\sigma^{-2}\right), \qquad (5)$$

and turn the inference into an optimization problem where $q(\theta)$ is obtained by minimizing its Kullback–Leibler divergence $\mathrm{KL}(q\|p)$ from the target distribution, $p(\theta \mid y)$. Because the marginal log-likelihood may be written as

$$\log p(y) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \qquad (6)$$

where

$$\mathcal{L}(q) = \int q(\theta) \log\left\{\frac{p(y,\theta)}{q(\theta)}\right\} \mathrm{d}\theta, \qquad \mathrm{KL}(q\|p) = -\int q(\theta) \log\left\{\frac{p(\theta \mid y)}{q(\theta)}\right\} \mathrm{d}\theta,$$

minimizing the Kullback–Leibler divergence amounts to maximizing $\mathcal{L}(q)$, which represents a lower bound for $\log p(y)$. To this end, we observe that

$$
\begin{aligned}
\mathcal{L}(q) &= \int \prod_{k=1}^{J} q_k(\theta_k) \left\{\log p(y,\theta) - \sum_{k=1}^{J} \log q_k(\theta_k)\right\} \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_J \\
&= \int q_j(\theta_j) \left\{\int \log p(y,\theta) \prod_{k \neq j} q_k(\theta_k) \mathrm{d}\theta_k - \log q_j(\theta_j)\right\} \mathrm{d}\theta_j + \mathrm{cst} \\
&= \int q_j(\theta_j) \log\left\{\frac{p_{-j}(\theta_j; y)}{q_j(\theta_j)}\right\} \mathrm{d}\theta_j + \mathrm{cst}, \qquad\qquad j = 1, \dots, J, \quad (7)
\end{aligned}
$$

where cst is constant with respect to θ_j and where we introduced the distribution

$$p_{-j}(\theta_j; y) = \mathrm{cst} \times \exp\left[\mathrm{E}_{-j}\left\{\log p(y,\theta)\right\}\right],$$

with $\mathrm{E}_{-j}\{\cdot\}$ denoting the expectation with respect to the distributions q_k over all variables θ_k, $k \neq j$. The right-hand side of (7) corresponds to the negative Kullback–Leibler divergence between $q_j(\theta_j)$ and $p_{-j}(\theta_j; y)$, plus a constant. Hence, assuming that the $q_k(\theta_k)$, $k \neq j$, are fixed, the

distribution $q_j(\theta_j)$ which maximizes $\mathcal{L}(q)$ is $q_j(\theta_j) = p_{-j}(\theta_j; y)$, i.e., the maximum of $\mathcal{L}(q)$ occurs when

$$\log q_j(\theta_j) = \mathrm{E}_{-j}\{\log p(y, \theta)\} + \mathrm{cst}, \qquad j = 1, \ldots, J. \tag{8}$$

The relations (8) give rise to cyclic dependencies among the densities $q_j(\theta_j)$. This suggests an iterative algorithm whose convergence can easily be monitored by evaluating changes in the lower bound $\mathcal{L}(q)$. Our choice (5) ensures that the coordinate updates can be derived in closed form; in particular, the semi-conjugacy of our model implies that the prior densities of all parameters are preserved by the variational densities. For instance, a spike-and-slab distribution with modified parameters is recovered at posterior level, $q(\beta_{st}, \gamma_{st}) = q(\beta_{st} \mid \gamma_{st}) q(\gamma_{st})$, with

$$\beta_{st} \mid \gamma_{st} = 1, y \sim \mathcal{N}\left(\mu_{\beta,st}, \sigma_{\beta,st}^2\right), \qquad \beta_{st} \mid \gamma_{st} = 0, y \sim \delta_0, \qquad \gamma_{st} \mid y \sim \mathrm{Bernoulli}\left(\gamma_{st}^{(1)}\right),$$

where the *variational parameters* $\mu_{\beta,st}$, $\sigma_{\beta,st}^2$, $\gamma_{st}^{(1)}$ are to be updated iteratively. Convergence is ensured by the convexity of $\mathcal{L}(q)$ in each of the $q_j(\theta_j)$ (Boyd and Vandenberghe, 2004, §§ 3.1.5, 3.2.4, 3.2.5). The algorithm and its derivation are given in Appendix B.

4 Empirical quality assessment of the variational approximation

4.1 Tightness of the marginal log-likelihood lower bound. In this section we evaluate the closeness of the variational density q to the target posterior distribution by approximating the Kullback–Leibler divergence $\mathrm{KL}(q\|p)$. Because of relation (6), this amounts to assessing the tightness of the variational lower bound for the marginal log-likelihood, $\mathcal{L}(q)$. For small problems, the likelihood $p(y)$ may be accurately approximated using simple Monte Carlo sums. We have

$$p(y) = \int \cdots \int \mathrm{d}\omega\, \mathrm{d}\sigma^{-2} \left\{\prod_{s=1}^{p} p(\omega_s)\right\} p\left(\sigma^{-2}\right) \prod_{t=1}^{d} \left\{\sum_{\gamma_t \in \{0,1\}^p} p\left(y_t \mid \gamma_t, \sigma^{-2}\right) \prod_{s=1}^{p} p\left(\gamma_{st} \mid \omega_s\right)\right\},$$

with

$$p\left(y_t \mid \gamma_t, \sigma^{-2}\right) = \begin{cases} (2\pi)^{-n/2}\, \Gamma\left(\dfrac{n}{2} + \eta_t\right) \dfrac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)} \left(\kappa_t + \dfrac{\|y_t\|^2}{2}\right)^{-n/2 - \eta_t}, & q_{\gamma_t} = 0, \\[4mm] (2\pi)^{-n/2} \left|V_{\gamma_t, \sigma^{-2}}\right|^{-1/2} \Gamma\left(\dfrac{n}{2} + \eta_t\right) \dfrac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)} \left(\kappa_t + \dfrac{S_{\gamma_t}^2}{2}\right)^{-n/2 - \eta_t} \left(\sigma^{-2}\right)^{q_{\gamma_t}/2}, & \\ & \text{otherwise}, \end{cases}$$

where

$$q_{\gamma_t} = \sum_{s=1}^{p} \gamma_{st}, \qquad V_{\gamma_t, \sigma^{-2}} = X_{\gamma_t}^T X_{\gamma_t} + \sigma^{-2} I_{q_{\gamma_t}}, \qquad S_{\gamma_t, \sigma^{-2}}^2 = \|y_t\|^2 - y_t^T X_{\gamma_t} V_{\gamma_t, \sigma^{-2}}^{-1} X_{\gamma_t}^T y_t;$$

see Appendix C for details. As no closed form is available for the remaining integrals, we use

$$p(y) \approx \frac{1}{I} \sum_{i=1}^{I} \prod_{t=1}^{d} \left\{\sum_{\gamma_t \in \{0,1\}^p} p\left(y_t \mid \gamma_t, \left(\sigma^{-2}\right)^{(i)}\right) \prod_{s=1}^{p} p\left(\gamma_{st} \mid \omega_s^{(i)}\right)\right\},$$

where we independently generate

$$\left(\sigma^{-2}\right)^{(i)} \sim \mathrm{Gamma}(\lambda, \nu), \qquad \omega_s^{(i)} \sim \mathrm{Beta}(a_s, b_s), \qquad s = 1, \ldots, p, \qquad i = 1, \ldots, I. \tag{9}$$

8

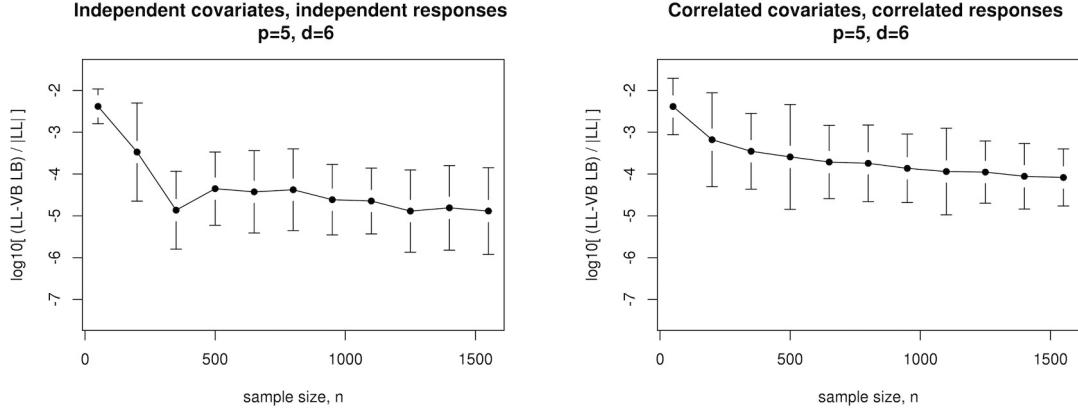Figure 2: Log_{10} relative difference between the marginal log-likelihood and the variational lower bound. Left: independent covariates and responses. Right: correlated covariates and responses, $\rho = 0.75$. Problems with $p = 5$ covariates, of which $p_0 = 3$ randomly selected as "active" (associated with at least one response), and $d = 6$ responses, of which $d_0 = 3$ "active" (associated with at least one covariate). Each active covariate is associated with an additional active response with probability 0.25 and explains on average 3.5% of the variance of its corresponding response(s). The number of draws for the simple Monte Carlo approximations is $I = 50,000$; the number of replicates for each sample size is 150.

Figure 2 displays the relative difference $\{\log p(y) - \mathcal{L}(q)\} / \log p(y)$ for problems with $p = 5$ covariates, $d = 6$ responses and increasing sample sizes, n. In the left panel, the covariates are independent of each other, and so are the responses. In the right panel, the covariates are equicorrelated with correlation coefficient $\rho = 0.75$, and so are the responses. In both cases, the mean relative difference is below 1% with $n = 50$ and seems to decrease as n grows. Although we are not aware of any such study with which to benchmark our results, these values seem very small, suggesting that our variational distribution q adequately reflects the target distribution p, at least for small problems. Likewise, the variational lower bound $\mathcal{L}(q)$ may be used as a proxy for the marginal log-likelihood when performing model selection; this use will be illustrated in Section 5.3. The fact that the variational lower bound remains tight in the correlated data case is reassuring, as it suggests that the independence assumptions underlying the mean-field factorization of q may only weakly impact the quality of the approximation.

4.2 Comparison with Markov Chain Monte Carlo. We complement our quality assessment by comparing several variational posterior quantities with those for MCMC inference on problems of moderate size. A fair comparison is not straightforward, as these two types of inference rely on stopping rules and convergence diagnostics of very different natures. While the convergence criterion for variational inference comes down to a tolerance to be prescribed, the ability of MCMC sampling to adequately explore the model space for a given chain length can be difficult to evaluate, and usually varies greatly with the problem size. To alleviate the risk of inaccurate MCMC inference, we run 10^5 iterations and discard the first half. We also support our comparison with selected quantities approximated by simple Monte Carlo sums, namely, the posterior probability

Table 2: Variational Bayes (VB), MCMC and simple Monte Carlo estimates for β and ω (components corresponding to noise averaged). Standard errors are in parentheses.

10×	$\beta_{1,2}$	$\beta_{2,1}$	Active $\beta_{3,2}$	$\beta_{4,1}$	$\beta_{4,2}$	Inactive β_{rest} (avg)
Truth	−1.75	2.87	2.37	3.73	−4.76	0.00
VB	−1.74 (0.01)	1.86 (0.01)	1.70 (0.02)	2.26 (0.01)	−3.48 (0.01)	0.02 (0.04)
MCMC	−1.74 (0.33)	1.86 (0.32)	1.69 (0.34)	2.26 (0.32)	−3.48 (0.34)	0.02 (0.14)

	ω_1	ω_2	Active ω_3	ω_4		Inactive ω_{rest} (avg)
True prop. of active resp.	0.2	0.2	0.2	0.4		0
VB	0.21 (0.14)	0.25 (0.15)	0.19 (0.14)	0.33 (0.17)		0.03 (0.06)
MCMC	0.23 (0.18)	0.26 (0.18)	0.21 (0.16)	0.35 (0.18)		0.04 (0.08)
Simple Monte Carlo	0.25	0.26	0.21	0.35		0.05

of inclusion of a covariate X_s for a response y_t,

$$
p(\gamma_{st} = 1 \mid y) \;=\; \frac{1}{p(y)} \frac{1}{I} \sum_{i=1}^{I} \left[\prod_{t' \neq t} \left\{ \sum_{\gamma_{t'} \in \{0,1\}^p} p\left(y_{t'} \mid \gamma_{t'}, \left(\sigma^{-2}\right)^{(i)} \right) \prod_{s'=1}^{p} p\left(\gamma_{s't'} \mid \omega_{s'}^{(i)} \right) \right\} \right.
$$

$$
\left. \times \left\{ \sum_{\gamma_t \in \{0,1\}^p:\ \gamma_{st}=1} p\left(y_t \mid \gamma_t, \left(\sigma^{-2}\right)^{(i)} \right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}^{(i)} \right) \right\} \right],
$$

and the posterior mean of ω_s, controlling the proportion of responses associated with covariate X_s,

$$
\mathrm{E}(\omega_s \mid y) \;=\; \frac{1}{p(y)} \frac{1}{I} \sum_{i=1}^{I} \omega_s^{(i)} \prod_{t=1}^{d} \left\{ \sum_{\gamma_t \in \{0,1\}^p} p\left(y_t \mid \gamma_t, \left(\sigma^{-2}\right)^{(i)} \right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}^{(i)} \right) \right\},
$$

with the samples $\left(\sigma^{-2}\right)^{(i)}$ and $\{\omega_s^{(i)}\}$ generated as in (9), with $I = 2 \times 10^5$ draws.

Table 2 reports the variational, MCMC and simple Monte Carlo estimates of β and ω for a problem with $p = 8$ covariates, $d = 5$ responses for $n = 250$ samples, and with each nonzero association explaining on average 13.5% of response variance. The estimates all agree closely. Those of the five active regression coefficients, $\beta_{1,2}$, $\beta_{2,1}$, $\beta_{3,2}$, $\beta_{4,1}$ and $\beta_{4,2}$, are significantly different from zero, unlike the average estimate of the inactive coefficients. A plot of the MCMC and variational posterior densities (the latter obtained in closed form), given in Figure 7 of Appendix C.2, shows that the posterior modes of the inactive coefficients are all zero. Moreover, in this case the variational distributions are usually solely made up of a clear spike at zero, whereas the MCMC histograms correspond roughly to a centered Gaussian distribution with average standard deviation 0.014. Table 2 also indicates a shrinkage effect for both variational and MCMC posterior means of the nonzero β compared to the true values. This is a consequence of the spike-and-slab prior but does not seem to hamper the detection of the association signals, since the posterior probabilities of inclusion of the true nonzero associations are concentrated around 1, while those corresponding to noise are usually much lower, whether obtained by MCMC, variational or simple Monte Carlo procedures; see Figure 8 of Appendix C.2. Finally, the estimates of $\{\omega_s\}$ in Table 2 provide a fair approximation to the actual proportion of responses associated with a given covariate.

Two additional numerical experiments comparing variational and MCMC posterior quantities are provided in Appendix C.2. One compares the estimates of ω and τ with the true values when

the data are generated from the model with $p = 100$ covariates and $d = 10$ responses. It also provides receiver operating characteristic (ROC) curves assessing the pairwise variable selection performance for both inference types. The other simulation gathers the observed values, y, and the estimated posterior means of $X\beta$ obtained by variational and MCMC procedures and an oracle. Both experiments indicate equivalent performance for MCMC and variational inferences.

5 Statistical performance

5.1 Predictor selection. The problems considered in Section 4.2 were small enough to allow accurate and tractable MCMC inference. In this section, we assess the performance of our approach on larger problems by comparing it to popular variable selection methods; i.e., with joint modelling of outcomes and covariates (elastic net for multivariate Gaussian responses), with joint modelling of covariates only (Bayesian multiple regression based on MCMC inference, "BAS", or variational inference, "varbvs"), or with fully marginal modelling (univariate ordinary least squares and "lmBF" Bayesian regressions). Complete descriptions and references are in Appendix D.2. The methods are compared by measuring their ability to detect the active covariates, i.e., to determine which covariates are associated with at least one response. For our variational approach, this task is achieved by ranking the posterior means of the ω_s, which control the proportion of responses associated with a given covariate.

Our data-generation design is based on generally accepted principles of population genetics. We simulate SNPs under Hardy–Weinberg equilibrium from a binomial distribution with probabilities corresponding to minor allele frequencies of common variants, chosen in the interval $(0.05, 0.5)$ uniformly at random, and we generate outcomes from Gaussian distributions with specific error variances. The dependence structure of these variables is either enforced block-wise with preselected auto- or equicorrelation coefficients or chosen to be that of real data. The labels of the active SNPs and outcomes are picked randomly, and each active SNP is associated to one (randomly selected) active outcome and to each of the remaining active outcomes with a prescribed probability; some outcomes are therefore under pleiotropic control. The proportions of outcome variance explained per SNP are simulated for all associations from a positively skewed Beta distribution to favour the generation of smaller effects, and they are then rescaled to match a given average proportion. To mimic the result of natural selection, the effect sizes are inversely related to the SNP minor allele frequencies. For more details, see Appendix D.1.

We perform 48 replications for each of three simulation configurations. The first configuration has moderate numbers of covariates ($p = 5,000$) and outcomes ($d = 50$), and allows time-consuming methods to run within hours. The second has many outcomes ($d = 20,000$) and the third has many covariates ($p = 150,000$); these numbers approach those encountered in molecular QTL studies. The remaining settings (numbers of active outcomes and covariates, of observations, effect sizes, etc) are detailed in the caption to Figure 3.

The ROC curves in Figure 3 indicate that our approach outperforms the other methods. It is appreciably more powerful for low false positive rates, which are of particular interest for the highly sparse scenarios typically expected for genome-wide association studies. Despite the correlation among the covariates and the outcomes, our method does not seem to suffer from the independence assumptions implied by the mean-field approximation, as suggested by the results of Section 4.1.
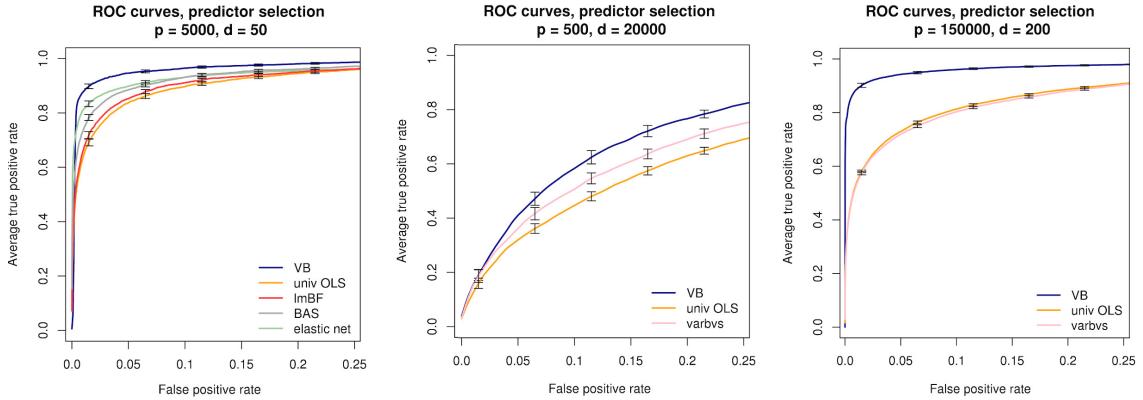
11

Figure 3: Truncated average receiver operating characteristic curves with 95% confidence intervals for predictor selection obtained from 48 replications. The competing methods are used in three studies of different sizes, based on their computational tractability. Left: $p = 5,000$ covariates spatially autocorrelated with correlation coefficient $\rho_X = 0.75$, $d = 50$ outcomes equicorrelated by blocks with four blocks of equal sizes and correlation coefficients $\rho_Y = 0.8, 0.3, 0.2$ and 0.5, $p_0 = 100$ active covariates, $d_0 = 40$ active outcomes, $n = 250$ observations, probability of association with an additional outcome $p_{\mathrm{add}} = 0.15$, average outcome variance percentage explained by the active covariates $p_{\mathrm{ve}} = 30.0\%$. Middle: $p = 500$ independent covariates, $d = 20,000$ outcomes equicorrelated by blocks of size 10 with $\rho_X \in \{0.5, \ldots, 0.8\}$, $p_0 = 300$, $d_0 = 12,500$, $n = 300$, $p_{\mathrm{add}} = 0.01$, $p_{\mathrm{ve}} = 55.8\%$. Right: $p = 150,000$ covariates autocorrelated by blocks of size 100 with $\rho_X \in \{0.5, \ldots, 0.9\}$, $d = 200$ outcomes with same correlation structure than real protein expression levels (Diogenes study, Larsen et al., 2010, see Appendix E.1), $p_0 = 500$, $d_0 = 150$, $n = 200$, $p_{\mathrm{add}} = 0.05$, $p_{\mathrm{ve}} = 62.6\%$. The univariate ordinary least squares and varbvs curves overlap.

The marginal ordinary least squares and marginal lmBF regressions appear to miss many associations because of their univariate modelling of covariates, but jointly accounting for the covariates may not suffice, as suggested by the rather poor performances of the Bayesian multiple regression approaches, BAS and varbvs, which apply separate multiple linear regressions for each outcome. It appears that the ability of our approach to exploit the similarity across outcomes yields more power to detect their shared associations. Finally, even though the multivariate elastic net models jointly the covariate and outcome variables, its inference suffers from the assumption that to each covariate corresponds a single regression coefficient, shared for all responses. As a consequence, regression estimates of covariates with weak or few associations with the responses may be shrunk to zero.

5.2 Combined selection of predictors and outcomes. Unlike the classical variable selection methods used as comparators in Section 5.1, our approach and those of Richardson et al. (2010), HESS, and Scott-Boyer et al. (2012), iBMQ, are tailored to molecular QTL problems: they quantify the associations between each covariate-response pair in a single model, and thus provide flexible and unified frameworks for detecting pairs of associated SNP-molecules, as well as pleiotropic SNPs associated with many molecular outcomes. In this section, we compare the three approaches in terms of the posterior quantities used to perform such selection. As both HESS and iBMQ rely on MCMC sampling, we consider smaller problems than in Section 5.1 in order to ensure convergence within a reasonable time. The simulated datasets have $p = 250$ covariates, of which $p_0 = 50$ are active, and $d = 100$ outcomes, of which $d_0 = 50$ are active, the probability of association being
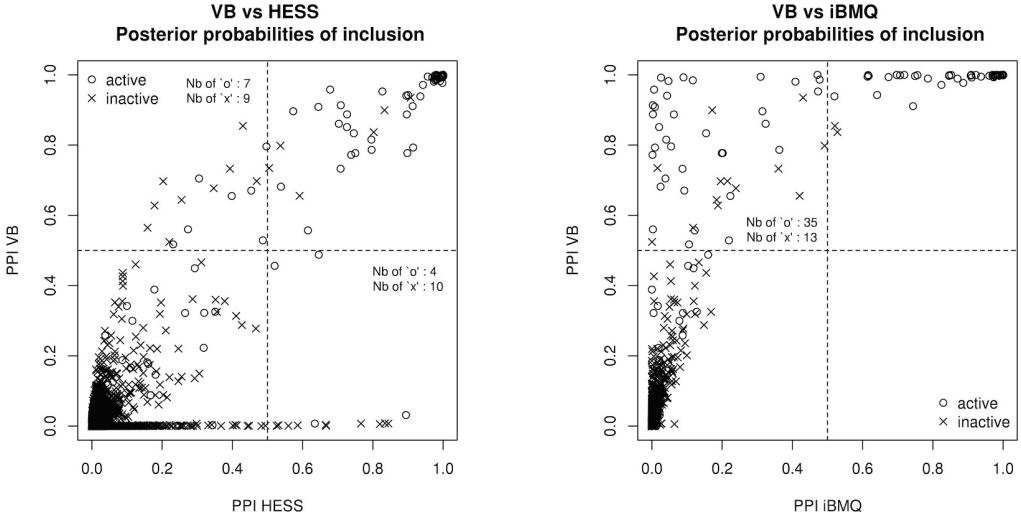
Figure 4: Marginal posterior probabilities of inclusion (PPI) obtained by our approach, and those of HESS (left) and those of iBMQ (right), for a problem with $p = 250$ covariates of which $p_0 = 50$ are active, with $d = 100$ outcomes, of which $d_0 = 50$ are active, and $n = 250$ samples.

Table 3: Mean true positive rate (TPR) and true negative rate (TNR) for our approach, HESS and iBMQ based on median probability models. Settings: $p = 250$, $p_0 = 50$, $d = 100$, $d_0 = 50$, $n = 250$, 48 replicates. Standard errors are in parentheses.

$100\times$	TPR	TNR
VB	58.9 (5.0)	99.9 (0.0)
HESS	57.9 (5.3)	99.9 (0.0)
iBMQ	0.1 (0.2)	99.8 (0.0)

0.05, for $n = 250$ samples. On average, the active covariates account for 22% of the variance of an outcome with which they are associated. HESS was run with three MCMC chains, the number selected by the authors for their simulations but with $50,000$ iterations of which $25,000$ were discarded as burn-in. For iBMQ, $50,000$ iterations were saved after removal of $50,000$ burn-in samples, as suggested in the package documentation for a problem of comparable dimensions. Inference for one replication took on average 10 seconds with our method, around 21 minutes with iBMQ and 4 hours with HESS (GPU computation option disabled, since no GPU was available to us) on an Intel Xeon CPU at 2.60 GHz with 64 GB RAM.

Figure 4 compares the marginal posterior probabilities of inclusion obtained by our method with those of HESS and iBMQ. We observe a strong correlation between our approach and HESS, with a quite good ability to discriminate between active and inactive covariate-response pairs. There is a discrepancy at the zero ordinate, where HESS signals a series of false positives and few true positives. The comparison with iBMQ is more contrasted, as the values of its posterior probabilities of inclusion for many true associations are below 0.1 and indistinguishable from noise. The same conclusions are reached when running the three methods on 47 additional datasets, as suggested by Table 3, which gathers sensitivity and specificity measures based on median probability models (Barbieri and Berger, 2004) (consisting of those covariate-response pairs whose posterior inclusion
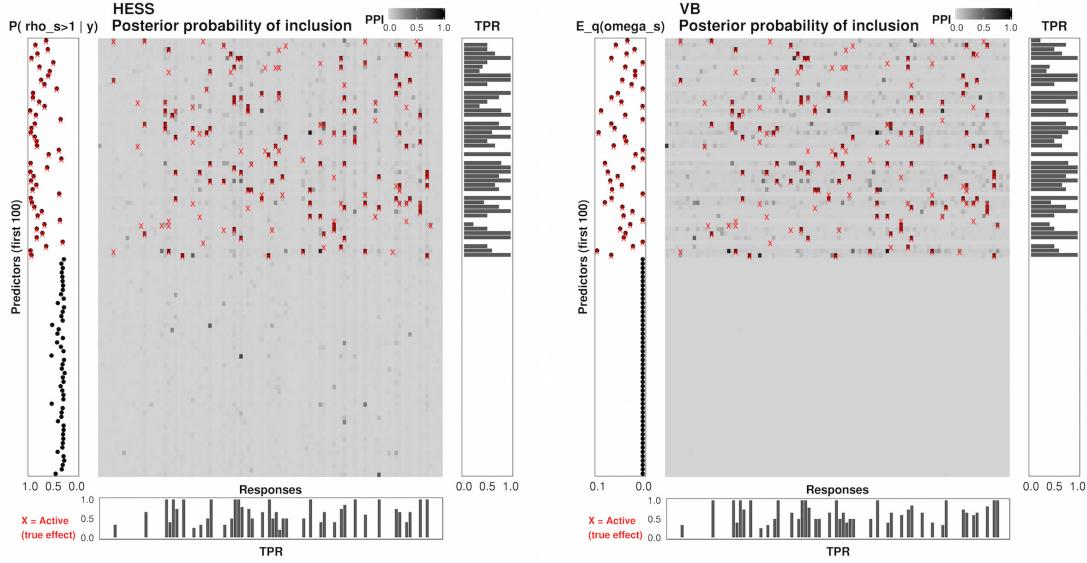
Figure 5: Posterior quantities for detection of associations with HESS (left) and with our approach (right), for a simulated dataset with $p = 250$ independent covariates (here only the first 100 are displayed), of which $p_0 = 50$ are active and placed first, with $d = 100$ responses (50 active) and $n = 250$ individuals. Marginal posterior probabilities of inclusion (central panel), true positive rates for predictor and response selection based on posterior probability of inclusion being > 0.5 (bottom and right panels), posterior probability $\mathrm{pr}(\rho_s > 1 \mid y)$ for HESS and posterior mean $\mathrm{E}_q(\omega_s \mid y)$ for our approach (left panel). The simulated associations are shown by red crosses.

probability is higher than 0.5). As discussed in Section 2, control of signal sparsity can be induced through the prior for ω_s, still, rather than median probability models, one may prefer to use a data-driven false discovery threshold in order to prescribe a desired level of false discoveries.

Figure 5 compares the patterns recovered by HESS and by our method, again based on the marginal posterior probabilities of inclusion from the first replicate. Visual comparison of the true positive rates suggests that the abilities of the two approaches to detect the true associations are very similar. Our approach indicates the presence of associations in the region of active covariates only, whereas the HESS pattern is blurrier in regions of inactive covariates. The posterior means of $\{\omega_s\}$ from our approach discriminate quite well between active and inactive covariates, and so do, for HESS, the posterior probabilities $\mathrm{pr}(\rho_s > 1 \mid y)$ $(s = 1, \ldots, p)$, described by Richardson et al. (2010) as capturing the propensity for a given covariate to influence several responses simultaneously.

5.3 Application to a real mQTL dataset. We end these numerical experiments by illustrating our approach on data from a large multicenter dietary intervention study called Diogenes (Larsen et al., 2010). The study contains a series of genomic data types collected at different stages of a dietary treatment provided to the cohort. Its goal is to uncover molecular mechanisms underlying the metabolic status of overweight individuals and improve understanding of the factors predisposing weight regain after a diet. Here, we perform a metabolite quantitative trait locus (mQTL) analysis; in this context, the metabolites may be viewed as proxies for the clinical condition of

14

Table 4: Number of associations declared by our method and by varbvs, and number of signals in common at selected permutation-based false discovery rates. For each case, the number of associations also declared by univariate screening at Benjamini–Hochberg FDR of 25% is in parentheses.

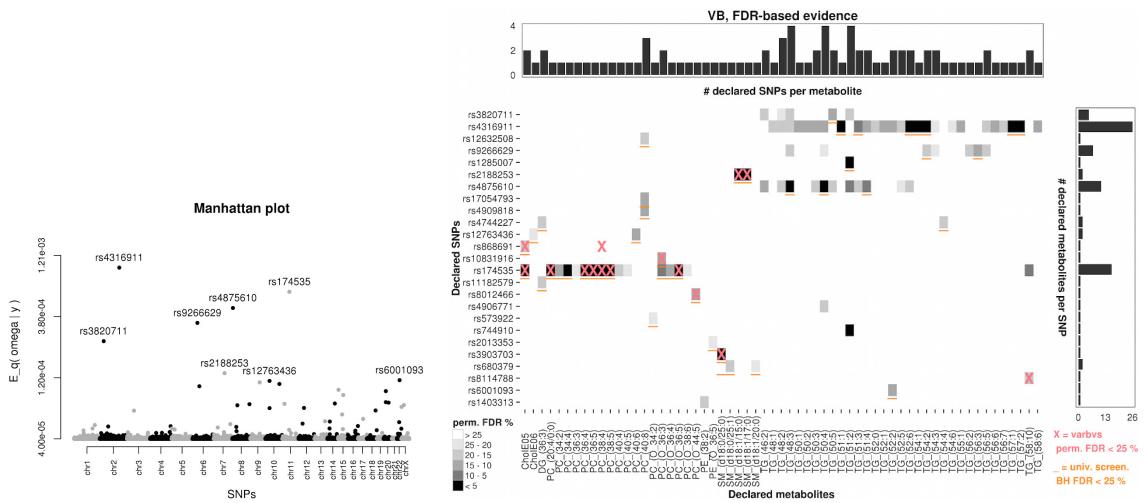| | # declared: | | |
Permutation-based FDR (%)	VB	varbvs	VB ∩ varbvs
5	21	19	8
10	26	19	8
15	47	21	10
20	76	31	12
25	89 (48 univ.)	47 (19 univ.)	14 (13 univ.)

Figure 6: SNPs and pairwise associations declared by our approach for the Diogenes study. Left: Manhattan plot for SNP association and evidence of pleiotropy. Right: posterior probabilities of inclusion for SNP-metabolite associations declared at estimated FDR of 25% and overlap with the associations declared by the varbvs method at same FDR level (pink crosses) and declared by univariate screening at Benjamini–Hochberg FDR of 25% (orange underscores).

interest, weight maintenance. We also use this illustration on real data to further highlight the benefits of modelling the outcomes jointly via an extensive permutation-based comparison with the single-response variational method varbvs (Carbonetto and Stephens, 2012).

After quality control, the data consist of $p = 215,907$ tag SNPs and $d = 125$ metabolite expression levels, adjusted for age, center and gender, for $n = 317$ individuals. The SNPs were collected on Illumina HumanCore arrays and the metabolites were quantified in plasma using liquid chromatography-mass spectrometry (LC-MS). They span cholesterol esters (CholE), phosphatidyl-cholines (PC), phosphatidylethanolamines (PE), sphingomyelins (SM), di- (DG) and triglycerides (TG). Appendix E.1 provides more details.

In order to adjust for multiplicity, we specify the hyperparameters for ω according to the discussion of Section 2 and choose the prior average number of active SNPs, p^*, by grid search within a 3-fold cross-validation procedure that maximizes the variational lower bound. After hyperparameter selection, the algorithm converged in 83 iterations, taking about 10 hours on an Intel Xeon CPU at 2.60 GHz with 512 GB RAM. The posterior means $\mathrm{E}_q(\omega_s \mid y)$ suggest

15

the presence of several active SNPs, spread across the chromosomes (Figure 6), but we use the marginal posterior inclusion probabilities, $\mathrm{E}_q(\gamma_{st} \mid y)$, to declare pairwise associations and active SNPs.

We compare varbvs and our method on real data based on the number of associations declared by each method at specific false discovery rates estimated by permutations. We apply Efron's Bayesian interpretation of the false discovery rate (Efron, 2008) to posterior probabilities of inclusion, and use an empirical null distribution based on $B = 400$ permutations to compute the estimate

$$\widehat{\mathrm{FDR}}(\tau) = \frac{\mathrm{median}_{b=1,\ldots,B} \#\{\mathrm{PPI}_{st}^{(b)} > \tau\}}{\#\{\mathrm{PPI}_{st} > \tau\}}, \qquad 0 < \tau < 1, \tag{10}$$

for a grid of thresholds τ; we then fit a cubic spline to the resulting false discovery rates to find thresholds for specific rates. The analysis suggests that our method is more powerful, with 89 associations declared at an estimated FDR of 25%, against 47 for varbvs; the superiority of our method is further highlighted by Table 4. Figure 6 displays the associations declared by our method and the overlap with those declared by varbvs at estimated FDR of 25%; the associations detected also largely agree with those obtained with marginal screening at Benjamini–Hochberg FDR of 25%.

Database searches on the functional relevance of the detected associations give hints of promising biological functions related to metabolic activities for 12 of the 25 SNPs declared as active by our procedure. For instance, the most outstanding SNP Figure 6, $rs4316911$, shows many associations with triglyceride levels, and turns out to be located less than $150kb$ from the protein coding gene ITGA6 known to be linked to diabetic kidney disease (Iyengar et al., 2015). The second most prominent pleiotropic SNP, $rs174535$, is declared by our approach to be associated with phospholipids, more precisely with 14 different phosphatidylcholine levels, of which four are ether-linked/plasmalogen (PC-O). Interestingly, this latter SNP has been recently reported to be related to metabolite levels; among others, it was found to be associated with trans fatty acid levels and plasma phospholipid levels (Mozaffarian et al., 2015), in line with our findings. Moreover, it was found to be an eQTL for the fatty acid desaturase genes FADS1 and FADS2. The SNP $rs3903703$ too has been identified as associated with very long-chain fatty acid levels (Lemaitre et al., 2015). This seems to agree with our findings, in which $rs3903703$ exhibits associations with sphingomyelin, a type of lipid containing fatty acids of different chain lengths. The complete subset of SNPs with metabolism-related links found by our procedure is given in Appendix E. Additional details on this real data study, as well as on its replication using simulated data, are also provided there.

6   Conclusions

We have described a scalable and efficient approach to joint variable selection from large numbers of candidate predictor and outcome variables. As it exploits the similarity across outcomes through a flexible hierarchical structure, our procedure outperforms the most popular predictor selection approaches in high-dimensional set-ups. The variational approximation on which our approach relies provides accurate posterior quantities, with reduced computational effort relative to MCMC procedures; in particular, it yields inferences comparable to those of the MCMC procedure HESS (Richardson et al., 2010). Convergence control is automatic, whereas convergence assess-

ment for MCMC algorithms can be difficult, especially in high dimensions. Our simulations also show that variable selection remains powerful when the predictors and outcomes are correlated, notwithstanding the independence assumptions underlying the mean-field factorization.

The key added-value of our approach is its applicability to molecular QTL datasets without the need for prior dimension reduction. In an application, our approach recovered several previously reported SNP-metabolite associations, and declared more associations than the single-outcome method "varbvs" (Carbonetto and Stephens, 2012) at prescribed false discovery rates, thus highlighting the benefits of jointly modelling the outcomes. To the best of our knowledge, no competing Bayesian approach for joint inference on two high-dimensional sets of variables can deal with the problem sizes typically encountered in molecular QTL analyses.

Bioinformatics is moving towards whole-genome analyses, for which several million genetic variants need to be considered, so it seems worthwhile to consider further speed-up strategies for our approach. One option is to use new optimization procedures. So-called natural gradient methods, which rely on the Riemannian structure of variational approximate distributions, seem particularly attractive, as they can be orders of magnitude faster than conventional gradient algorithms (Honkela et al., 2008). Another possibility is a "split-and-merge" strategy, i.e., first partitioning the variable space and then inferring a global variational distribution on the aggregated dataset. Tran et al. (2016) designed a variant of this approach for sample space partitioning. At the recombination step, they proposed to "merge" the variational distributions by exploiting the independence assumptions of the mean-field formulation. Both strategies could lead to significant computational gains.

7 Software

The algorithm and the data-generation functions used in this paper are implemented in the publicly available R package `locus`.

Acknowledgements

Funding

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, Inc., New York, 1965.

H. Attias. A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems*, 12:209–215, 2000.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32: 870–897, 2004.

D. Bates and M. Maechler. *Matrix: Sparse and dense matrix classes and methods, 2015*, 2015. R package version 1.2.2.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. De Haan, A. I. Su, et al. Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics*, 4:e1000232, 2008.

P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7:73–108, 2012.

M. Civelek and A. J. Lusis. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, 15:34–48, 2014.

M. Clyde. *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*, 2016. R package version 1.0.9.

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23:1–22, 2008.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with Discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

T. Flutre, X. Wen, J. Pritchard, and M. Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, 9:e1003486, 2013.

J. H. Friedman, T. J. Hastie, and R. J. Tibshirani. *glmnet: Lasso and elastic-net regularized generalized linear models*, 2009. R package version 2.0.2.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5:1780–1815, 2011.

N. J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22:329–343, 2002.

A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In M. Ishikawa, Kenji Doya, H. Miyamoto, and T. Yamakawa, editors, *Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part II*, pages 305–314. Springer, Berlin, 2008.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33:730–773, 2005.

S. K. Iyengar, J. R. Sedor, B. I. Freedman, W. L. Kao, M. Kretzler, B. J. Keller, H. E. Abboud, S. G. Adler, L. G. Best, and D. W. Bowden. Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: family investigation of nephropathy and diabetes (FIND). *PLoS Genetics*, 11:e1005352, 2015.

Z. Jia and S. Xu. Mapping quantitative trait loci for expression abundance. *Genetics*, 176:611–623, 2007.

L. Jiang, J. Liu, X. Zhu, M. Ye, L. Sun, X. Lacaze, and R. Wu. 2HiGWAS: a unifying high-dimensional platform to infer the global genetic architecture of trait development. *Briefings in Bioinformatics*, 16:bbv002, 2015.

D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, and D. J. Thomas. The UCSC genome browser database. *Nucleic Acids Research*, 31:51–54, 2003.

T. M. Larsen, S.-M. Dalskov, M. van Baak, S. A. Jebb, A. Kafatos, A. F. H. Pfeiffer, J. A. Martinez, T. Handjieva-Darlenska, M. Kunesova, C. Holst, W. H. M. Saris, and A. Astrup. The Diet, Obesity and Genes (Diogenes) Dietary study in eight European countries—a comprehensive design for long-term intervention. *Obesity Reviews*, 11:76–91, 2010.

R. N. Lemaitre, I. B. King, E. K. Kabagambe, J. H. Y. Wu, B. McKnight, A. Manichaikul, W. Guan, Q. Sun, D. I. Chasman, and M. Foy. Genetic loci associated with circulating levels of very long-chain saturated fatty acids. *Journal of Lipid Research*, 56:176–184, 2015.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, and N. Young. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45:580–585, 2013.

R. D. Morey and J. N. Rouder. *BayesFactor: computation of Bayes factors for common designs*, 2015. R package version 0.9.12-2.

D. Mozaffarian, E. K. Kabagambe, C. O. Johnson, R. N. Lemaitre, A. Manichaikul, Q. Sun, M. Foy, L. Wang, H. Wiener, and M. R. Irvin. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *The American Journal of Clinical Nutrition*, 101:398–406, 2015.

P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, 7: e34861, 2012.

J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64:140–153, 2010.

J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108:18026–18031, 2011.

M. Plummer, N. Best, K. Cowles, and K. Vines. *CODA: Convergence Diagnosis and Output Analysis for MCMC*, 2006. R package version 0.18.1.

M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14: 656–664, 1998.

S. Richardson, L. Bottolo, and J. S. Rosenthal. Bayesian models for sparse regression analysis of high-dimensional data. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, volume 9, pages 539–569. Oxford University Press, New York, 2010.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38:2587–2619, 2010.

M. P. Scott-Boyer, G. C. Imholte, A. Tayeb, A. Labbe, C. F. Deschepper, and R. Gottardo. An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11:1515–1544, 2012.

S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89:607–618, 2011.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14:483–495, 2013.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *OpenBUGS user manual, version 3.0. 2*, 2007. MRC Biostatistics Unit, Cambridge.

M.-N. Tran, D. J. Nott, A. Y. C. Kuk, and R. Kohn. Parallel variational Bayes for large datasets with an application to generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 25:626–646, 2016.

J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5:99–114, 1949.

N. Wang, K. Gosik, R. Li, B. Lindsay, and R. Wu. A block mixture model to map eQTLs for gene clustering and networking. *Scientific Reports*, 6:21193, 2016.

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, and L. Hindorff. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42:D1001–D1006, 2014.

E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean-field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 583–591, San Francisco, 2002. Morgan Kaufmann.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Studies in Bayesian Econometrics*, volume 6, pages 233–243. Elsevier, New York, 1986. P. K. Goel and A. Zellner, editors.

X. Zhou and M. Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature Methods*, 11:407, 2014.

A Multiplicity control at covariate level

Sparsity control at covariate level can be induced through the prior distribution of ω, by carefully selecting its hyperparameters. The prior probability that X_s is "active" (i.e., associated with at least one response) is

$$p\left(\cup_{t=1}^{d}\{\gamma_{st}=1\}\right) = \int \left\{1 - \prod_{t=1}^{d} p\left(\gamma_{st}=0 \mid \omega_s\right)\right\} p(\omega_s)\mathrm{d}\omega_s = 1 - \frac{\text{Beta}(a_s, b_s + d)}{\text{Beta}(a_s, b_s)},$$

and this after a little algebra equals

$$1 - \frac{\prod_{j=1}^{d}(b_s + d - j)}{\prod_{j=1}^{d}(a_s + b_s + d - j)},$$

so assuming exchangeability and setting

$$a_s \equiv 1, \qquad b_s \equiv d(p - p^*)/p^*, \qquad 0 < p^* < p, \tag{11}$$

implies that

$$p\left(\cup_{t=1}^{d}\{\gamma_{st}=1\}\right) = \frac{d}{b_s + d} = \frac{p^*}{p},$$

where p^* is interpreted as a prior average number of active covariates. The choice (11) yields a multiplicity adjustment as suggested by a plot (Figure 1) of the prior odds ratios, indicating the penalty induced by the prior when moving from $q_s - 1$ to q_s responses associated with X_s,

$$
\begin{aligned}
\text{POR}(q_s - 1 : q_s) &= \frac{p\left(\sum_{t=1}^{d}\gamma_{st} = q_s - 1\right)}{p\left(\sum_{t=1}^{d}\gamma_{st} = q_s\right)} = \frac{\text{Beta}(a_s + q_s - 1, b_s + d - q_s + 1)}{\text{Beta}(a_s + q_s, b_s + d - q_s)} \\
&= \frac{b_s + d - q_s}{a_s + q_s - 1}, \qquad\qquad\qquad q_s = 1, \ldots, d.
\end{aligned}
$$

B Derivation of the variational algorithm

B.1 Variational distributions. We provide the detailed derivation of our variational algorithm, which is given in Appendix B.3. We have

$$
\begin{aligned}
p\left(y, \beta, \gamma, \tau, \sigma^{-2}, \omega\right) &= p\left(y \mid \beta, \tau\right) p\left(\beta \mid \gamma, \tau, \sigma^{-2}\right) p\left(\gamma \mid \omega\right) p(\omega) p(\tau) p(\sigma^{-2}) \\
&= \left\{\prod_{t=1}^{d} p\left(y_t \mid \beta_t, \tau_t\right)\right\} \left\{\prod_{s=1}^{p}\prod_{t=1}^{d} p\left(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2}\right)\right\} \left\{\prod_{s=1}^{p}\prod_{t=1}^{d} p\left(\gamma_{st} \mid \omega_s\right)\right\} \\
&\quad \times \left\{\prod_{s=1}^{p} p(\omega_s)\right\} \left\{\prod_{t=1}^{d} p(\tau_t)\right\} p\left(\sigma^{-2}\right), \tag{12}
\end{aligned}
$$

where

$$p(y_t \mid \beta_t, \tau_t) = (2\pi)^{-n/2} (\tau_t)^{n/2} \exp\left(-\frac{\tau_t}{2} \|y_t - X\beta_t\|^2\right), \qquad y_t \in \mathbb{R},$$

$$p(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2}) = \left\{(2\pi)^{-1/2} (\sigma^{-2}\tau_t)^{1/2} \exp\left(-\frac{\sigma^{-2}\tau_t}{2}\beta_{st}^2\right)\right\}^{\gamma_{st}} \delta_0(\beta_{st})^{1-\gamma_{st}}, \qquad \beta_{st} \in \mathbb{R},$$

$$p(\gamma_{st} \mid \omega_s) = \omega_s^{\gamma_{st}} (1-\omega_s)^{1-\gamma_{st}}, \qquad \gamma_{st} = 0,1,$$

$$p(\omega_s) = \omega_s^{a_s-1}(1-\omega_s)^{b_s-1}/\mathrm{B}(a_s, b_s), \qquad 0 < \omega_s < 1,$$

$$p(\tau_t) = \tau_t^{\eta_t-1} \exp\left(-\kappa_t \tau_t\right) \kappa_t^{\eta_t}/\Gamma(\eta_t), \qquad \tau_t > 0,$$

$$p(\sigma^{-2}) = \left(\sigma^{-2}\right)^{\lambda-1} \exp\left(-\nu\sigma^{-2}\right) \nu^\lambda/\Gamma(\lambda), \qquad \sigma^2 > 0.$$

Let $\theta = \left(\beta, \gamma, \tau, \sigma^{-2}, \omega\right)$ and consider the following *mean-field* form for the variational approximation,

$$q(\theta) = \left\{\prod_{s=1}^{p} \prod_{t=1}^{d} q(\beta_{st}, \gamma_{st})\right\} \left\{\prod_{s=1}^{p} q(\omega_s)\right\} \left\{\prod_{t=1}^{d} q(\tau_t)\right\} q\left(\sigma^{-2}\right).$$

We obtain each component of this factorization using the formula

$$\log q_j(\theta_j) = \mathrm{E}_{-j}\{\log p(y, \theta)\} + \mathrm{cst}, \qquad j = 1, \ldots, J,$$

with $p(y, \theta)$ given in (12), where E_{-j} is the expectation with respect to the distribution q_k over all variables θ_k $(k \neq j)$ and where cst is a constant with respect to θ_j. Writing $\theta_j^{(r)}$ for the rth moment with respect to the approximate posterior distribution q_j of θ_j, we have

$$
\begin{aligned}
\log q(\beta_{st}, \gamma_{st}) &= \sum_{k=1}^{d} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\{\log p(y_k \mid \beta_k, \tau_k)\} + \sum_{j=1}^{p}\sum_{k=1}^{d} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\{\log p(\beta_{jk} \mid \gamma_{jk}, \tau_k, \sigma^{-2})\} \\
&\quad + \sum_{j=1}^{p}\sum_{k=1}^{d} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\{\log p(\gamma_{jk} \mid \omega_j)\} + \mathrm{cst} \\
&= -\frac{1}{2}\tau_t^{(1)}\mathrm{E}_{-\beta_{st}}\|y_t - X\beta_t\|^2 + \gamma_{st}\frac{1}{2}\left\{\mathrm{E}\left(\log \sigma^{-2}\right) + \mathrm{E}\left(\log \tau_t\right) - \log(2\pi)\right\} \\
&\quad -\gamma_{st}\frac{\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}}{2}\beta_{st}^2 + (1-\gamma_{st})\delta_0(\beta_{st}) + \gamma_{st}\mathrm{E}\left(\log\omega_s\right) + (1-\gamma_{st})\mathrm{E}\{\log(1-\omega_s)\} + \mathrm{cst} \\
&= -\gamma_{st}\frac{\tau_t^{(1)}}{2}\left[\beta_{st}^2\left\{\|X_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\} - 2\beta_{st}\left\{y_t^T X_s - X_s^T \sum_{j=1, j\neq s}^{p} \beta_{jt}^{(1)} X_j\right\}\right] \\
&\quad + \gamma_{st}\frac{1}{2}\left\{\mathrm{E}\left(\log\sigma^{-2}\right) + \mathrm{E}\left(\log\tau_t\right) - \log(2\pi)\right\} + (1-\gamma_{st})\delta_0(\beta_{st}) + \gamma_{st}\mathrm{E}\left(\log\omega_s\right) \\
&\quad + (1-\gamma_{st})\mathrm{E}\{\log(1-\omega_s)\} + \mathrm{cst},
\end{aligned}
$$

where cst is a constant with respect to β_{st} and γ_{st}. Completing the square yields

$$
\begin{aligned}
q(\beta_{st}, \gamma_{st}) &= \mathrm{cst}\left[\left(2\pi\sigma_{\beta,st}^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma_{\beta,st}^2}(\beta_{st} - \mu_{\beta,st})^2\right\}\right]^{\gamma_{st}} \\
&\quad \times \left[\left[\exp\left\{\mathrm{E}\left(\log\sigma^{-2}\right) + \mathrm{E}\left(\log\tau_t\right)\right\}\sigma_{\beta,st}^2\right]^{1/2}\exp\left(\frac{1}{2}\mu_{\beta,st}^2\sigma_{\beta,st}^{-2}\right)\exp\{\mathrm{E}\left(\log\omega_s\right)\}\right]^{\gamma_{st}} \\
&\quad \times \{\delta_0(\beta_{st})\}^{1-\gamma_{st}}\exp\left[\mathrm{E}\{\log(1-\omega_s)\}\right]^{1-\gamma_{st}},
\end{aligned}
$$

with

$$\mu_{\beta,st} = \sigma_{\beta,st}^2 \tau_t^{(1)}\left\{y_t^T X_s - X_s^T \sum_{j=1, j\neq s}^{p} \mu_{\beta,jt}\gamma_{jt}^{(1)} X_j\right\}, \qquad \sigma_{\beta,st}^2 = \frac{1}{\tau_t^{(1)}\left\{\|X_s\|^2 + (\sigma^{-2})^{(1)}\right\}}.$$

We therefore observe that

$$q(\beta_{st}, \gamma_{st}) = q(\beta_{st} \mid \gamma_{st})q(\gamma_{st}),$$

with

$$\beta_{st} \mid \gamma_{st} = 1, y \sim \mathcal{N}\left(\mu_{\beta,st}, \sigma_{\beta,st}^2\right), \quad \beta_{st} \mid \gamma_{st} = 0, y \sim \delta_0, \quad \gamma_{st} \mid y \sim \text{Bernoulli}\left(\gamma_{st}^{(1)}\right),$$

and with

$$\frac{\gamma_{st}^{(1)}}{1 - \gamma_{st}^{(1)}} = \sigma_{\beta,st} \exp\left[\mathrm{E}\left(\log \omega_s\right) - \mathrm{E}\left\{\log\left(1 - \omega_s\right)\right\} + \frac{1}{2}\left\{\mathrm{E}\left(\log \tau_t\right) + \mathrm{E}\left(\log \sigma^{-2}\right)\right\} + \frac{1}{2}\mu_{\beta,st}^2 \sigma_{\beta,st}^{-2}\right],$$

i.e.,

$$\gamma_{st}^{(1)} = \left[1 + \sigma_{\beta,st}^{-1} \exp\left\{\mathrm{E}\left\{\log\left(1 - \omega_s\right)\right\} - \mathrm{E}\left(\log \omega_s\right) - \frac{1}{2}\mathrm{E}\left(\log \tau_t\right) - \frac{1}{2}\mathrm{E}\left(\log \sigma^{-2}\right) - \frac{1}{2}\mu_{\beta,st}^2 \sigma_{\beta,st}^{-2}\right\}\right]^{-1}. \tag{13}$$

We now compute the variational approximate distribution for the error variance of each y_t:

$$
\begin{aligned}
\log q(\tau_t) &= \mathrm{E}_{-\tau_t}\left\{\log p\left(y_t \mid \beta_t, \tau_t\right)\right\} + \sum_{s=1}^{p}\mathrm{E}_{-\tau_t}\left\{\log p\left(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2}\right)\right\} + \log p(\tau_t) + \text{cst} \\
&= \frac{n}{2}\log \tau_t - \frac{\tau_t}{2}\mathrm{E}\left(\|y_t - X\beta_t\|^2\right) + \frac{1}{2}\log \tau_t \sum_{s=1}^{p}\gamma_{st}^{(1)} - \frac{\tau_t}{2}\left(\sigma^{-2}\right)^{(1)}\sum_{s=1}^{p}\beta_{st}^{(2)} + (\eta_t - 1)\log \tau_t \\
&\quad -\kappa_t\tau_t + \text{cst} \\
&= \log \tau_t\left(\eta_t + \frac{n}{2} + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)} - 1\right) - \tau_t\left[\kappa_t + \frac{1}{2}\|y_t\|^2 - y_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}X_s\right. \\
&\quad \left.+ \sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}X_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}X_j + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\left\{\|X_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\}\right] + \text{cst}.
\end{aligned}
$$

Therefore we have

$$\tau_t \mid y \sim \text{Gamma}\left(\eta_t^*, \kappa_t^*\right)$$

and

$$\tau_t^{(1)} = \eta_t^* / \kappa_t^*,$$

where

$$
\begin{aligned}
\eta_t^* &= \eta_t + \frac{n}{2} + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}, \\
\kappa_t^* &= \kappa_t + \frac{1}{2}\|y_t\|^2 - y_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}X_s + \sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}X_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}X_j \\
&\quad + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\left\{\|X_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\}.
\end{aligned}
$$

Since τ_t has a Gamma distribution, the expectation $\mathrm{E}\left(\log \tau_t\right)$ appearing in (13) can be rewritten in terms of η_t^* and κ_t^* using the digamma function (Abramowitz and Stegun, 1965),

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)},$$

as

$$\mathrm{E}_q\left(\log \tau_t\right) = \Psi(\eta_t^*) - \log(\kappa_t^*). \tag{14}$$

23

We also find that

$$
\begin{aligned}
\log q\left(\sigma^{-2}\right) &= \sum_{s=1}^{p}\sum_{t=1}^{d} \mathrm{E}_{-\sigma^{-2}}\left\{\log p\left(\beta_{st}\mid\gamma_{st},\tau_t,\sigma^{-2}\right)\right\} + \log p(\sigma^{-2}) + \mathrm{cst} \\
&= \left(\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{d}\gamma_{st}^{(1)}\right)\log\sigma^{-2} - \sigma^{-2}\sum_{t=1}^{d}\frac{\tau_t^{(1)}}{2}\sum_{s=1}^{p}\beta_{st}^{(2)} + (\lambda-1)\log\sigma^{-2} - \nu\sigma^{-2} + \mathrm{cst} \\
&= \left(\lambda+\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{d}\gamma_{st}^{(1)}-1\right)\log\sigma^{-2} - \sigma^{-2}\left\{\nu+\sum_{t=1}^{d}\frac{\tau_t^{(1)}}{2}\sum_{s=1}^{p}\left(\sigma_{\beta,st}^2+\mu_{\beta,st}^2\right)\gamma_{st}^{(1)}\right\} + \mathrm{cst}\,.
\end{aligned}
$$

Thus

$$
\sigma^{-2}\mid y \sim \mathrm{Gamma}\left(\lambda^*,\nu^*\right)\,,
$$

$$
\left(\sigma^{-2}\right)^{(1)} = \lambda^*/\nu^*\,,
$$

where

$$
\lambda^* = \lambda + \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{d}\gamma_{st}^{(1)}\,, \qquad \nu^* = \nu + \frac{1}{2}\sum_{t=1}^{d}\tau_t^{(1)}\sum_{s=1}^{p}\left(\sigma_{\beta,st}^2+\mu_{\beta,st}^2\right)\gamma_{st}^{(1)}\,,
$$

and, as before, we now have

$$
\mathrm{E}_q\left(\log\sigma^{-2}\right) = \Psi(\lambda^*) - \log\nu^*\,. \tag{15}
$$

Finally, we have

$$
\begin{aligned}
\log q(\omega_s) &= \sum_{t=1}^{d}\mathrm{E}_{\gamma_{st}}\left\{\log p(\gamma_{st}\mid\omega_s)\right\} + \log p(\omega_s) + \mathrm{cst} \\
&= \left(\sum_{t=1}^{d}\gamma_{st}^{(1)}\right)\log\omega_s + \left\{\sum_{t=1}^{d}\left(1-\gamma_{st}^{(1)}\right)\right\}\log(1-\omega_s) + (a_s-1)\log\omega_s + (b_s-1)\log(1-\omega_s) + \mathrm{cst} \\
&= \left(a_s+\sum_{t=1}^{d}\gamma_{st}^{(1)}-1\right)\log\omega_s + \left(b_s-\sum_{t=1}^{d}\gamma_{st}^{(1)}+d-1\right)\log(1-\omega_s) + \mathrm{cst}\,,
\end{aligned}
$$

that is,

$$
\omega_s\mid y \sim \mathrm{Beta}\left(a_s^*,b_s^*\right)\,,
$$

and

$$
\omega_s^{(1)} = \frac{a_s^*}{a_s^*+b_s^*}\,,
$$

where

$$
a_s^* = a_s + \sum_{t=1}^{d}\gamma_{st}^{(1)} \qquad\qquad b_s^* = b_s - \sum_{t=1}^{d}\gamma_{st}^{(1)} + d\,.
$$

As ω_s has a Beta distribution, we also get (Abramowitz and Stegun, 1965)

$$
\begin{aligned}
\mathrm{E}_q\left(\log\omega_s\right) &= \Psi(a_s^*) - \Psi(a_s^*+b_s^*) = \Psi\left(a_s+\sum_{t=1}^{d}\gamma_{st}^{(1)}\right) - \Psi\left(a_s+b_s+d\right)\,, \\
\mathrm{E}_q\left\{\log(1-\omega_s)\right\} &= \Psi(b_s^*) - \Psi(a_s^*+b_s^*) = \Psi\left(b_s-\sum_{t=1}^{d}\gamma_{st}^{(1)}+d\right) - \Psi\left(a_s+b_s+d\right)\,. \tag{16}
\end{aligned}
$$

B.2 Lower bound of the marginal log-likelihood. We provide the computational details for the lower bound, $\mathcal{L}(q)$, of the marginal log-likelihood, $\log p(y)$. It is evaluated at each iteration of our algorithm, in order to monitor its convergence:

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(z) \log \left\{ \frac{p(y,z)}{q(z)} \right\} \mathrm{d}z \\
&= \sum_{t=1}^{d} A\left(y_t \mid \beta_t, \tau_t\right) + \sum_{s=1}^{p}\sum_{t=1}^{d} B\left(\beta_{st}, \gamma_{st} \mid \tau_t, \sigma^{-2}\right) + \sum_{t=1}^{d} C(\tau_t) + D\left(\sigma^{-2}\right) + \sum_{s=1}^{p} G(\omega_s),
\end{aligned}
$$

with

$$
\begin{aligned}
A\left(y_t \mid \beta_t, \tau_t\right) &= \mathrm{E}_q\left\{\log p(y_t \mid \beta_t, \tau_t)\right\} = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}\left(\log \tau_t\right) - \frac{1}{2}\tau_t^{(1)}\mathrm{E}\left\{\|y_t - X\beta_t\|^2\right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}\left(\log \tau_t\right) - \frac{1}{2}\tau_t^{(1)}\left\{ \|y_t\|^2 - 2y_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}X_s \right. \\
&\quad \left. +2\sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}X_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}X_j + \sum_{s=1}^{p}\|X_s\|^2\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\gamma_{st}^{(1)} \right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}\left(\log \tau_t\right) - \tau_t^{(1)}\left\{ \kappa_t^* - \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\left(\sigma^{-2}\right)^{(1)} - \kappa_t \right\},
\end{aligned}
$$

$$
\begin{aligned}
B\left(\beta_{st}, \gamma_{st} \mid \tau_t, \sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2})\right\} + \mathrm{E}_q\left\{\log p(\gamma_{st} \mid \omega_s)\right\} - \mathrm{E}_q\left\{\log q(\beta_{st}, \gamma_{st})\right\} \\
&= \frac{1}{2}\gamma_{st}^{(1)}\left\{ -\log(2\pi) + \mathrm{E}\left(\log \sigma^{-2}\right) + \mathrm{E}\left(\log \tau_t\right) - \left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right) \right\} \\
&\quad +\mathrm{E}_q\left\{(1 - \gamma_{st})\,\delta_0(\beta_{st})\right\} + \gamma_{st}^{(1)}\mathrm{E}\left(\log \omega_s\right) + \left(1 - \gamma_{st}^{(1)}\right)\mathrm{E}\left\{\log(1 - \omega_s)\right\} \\
&\quad +\frac{1}{2}\gamma_{st}^{(1)}\left\{\log(2\pi) + \log \sigma_{\beta,st}^2\right\} + \frac{1}{2\sigma_{\beta,st}^2}\mathrm{E}_q\left\{\gamma_{st}(\beta_{st} - \mu_{\beta,st})^2\right\} - \gamma_{st}^{(1)}\log \gamma_{st}^{(1)} \\
&\quad -\mathrm{E}_q\left\{(1 - \gamma_{st})\,\delta_0(\beta_{st})\right\} - \left(1 - \gamma_{st}^{(1)}\right)\log\left(1 - \gamma_{st}^{(1)}\right) \\
&= \frac{1}{2}\gamma_{st}^{(1)}\left\{\mathrm{E}\left(\log \sigma^{-2}\right) + \mathrm{E}\left(\log \tau_t\right)\right\} - \frac{1}{2}\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right) \\
&\quad +\gamma_{st}^{(1)}\mathrm{E}\left(\log \omega_s\right) + \left(1 - \gamma_{st}^{(1)}\right)\mathrm{E}\left\{\log(1 - \omega_s)\right\} + \frac{1}{2}\gamma_{st}^{(1)}\left(\log \sigma_{\beta,st}^2 + 1\right) \\
&\quad -\gamma_{st}^{(1)}\log \gamma_{st}^{(1)} - \left(1 - \gamma_{st}^{(1)}\right)\log\left(1 - \gamma_{st}^{(1)}\right),
\end{aligned}
$$

$$
\begin{aligned}
C(\tau_t) &= \mathrm{E}_q\left\{\log p(\tau_t)\right\} - \mathrm{E}_q\left\{\log q(\tau_t)\right\} \\
&= (\eta_t - \eta_t^*)\,\mathrm{E}\left(\log \tau_t\right) - (\kappa_t - \kappa_t^*)\,\tau_t^{(1)} + \eta_t\log \kappa_t - \eta_t^*\log \kappa_t^* - \log \Gamma(\eta_t) + \log \Gamma(\eta_t^*),
\end{aligned}
$$

$$
\begin{aligned}
D\left(\sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p\left(\sigma^{-2}\right)\right\} - \mathrm{E}_q\left\{\log q\left(\sigma^{-2}\right)\right\} \\
&= (\lambda - \lambda^*)\,\mathrm{E}\left(\log \sigma^{-2}\right) - (\nu - \nu^*)\left(\sigma^{-2}\right)^{(1)} + \lambda\log \nu - \lambda^*\log \nu^* - \log \Gamma(\lambda) + \log \Gamma(\lambda^*),
\end{aligned}
$$

$$
\begin{aligned}
G(\omega_s) &= \mathrm{E}_q\left\{\log p(\omega_s)\right\} - \mathrm{E}_q\left\{\log q(\omega_s)\right\} \\
&= (a_s - a_s^*)\,\mathrm{E}\left(\log \omega_s\right) + (b_s - b_s^*)\,\mathrm{E}\left\{\log(1 - \omega_s)\right\} - \log B(a_s, b_s) + \log B\left(a_s^*, b_s^*\right),
\end{aligned}
$$

where $\mathrm{E}_q\left(\log \tau_t\right)$, $\mathrm{E}_q\left(\log \sigma^{-2}\right)$, $\mathrm{E}_q\left(\log \omega_s\right)$ and $\mathrm{E}_q\left\{\log(1 - \omega_s)\right\}$ are given by (14), (15) and (16).

B.3 Variational algorithm.

Algorithm 1

inputs:

  y (centered), X (standardized using the usual unbiased estimator of the variance),

  a, b, η, κ, λ, ν, tol, maxit

initialize:

  $M = \{\mu_{\beta,st}\}$, $\Sigma = \{\sigma_{\beta,st}^2\}$, $\Gamma^{(1)} = \left\{\gamma_{st}^{(1)}\right\}$, $\tau^{(1)} = \left\{\tau_t^{(1)}\right\}$

  $\mathcal{L}(q) \leftarrow -\infty$, it $\leftarrow 0$

repeat:

  $\left(\sigma^{-2}\right)^{(1)} \leftarrow \lambda^*/\nu^*$,   where $\lambda^* = \lambda + \frac{1}{2}\mathbb{1}_p^T\,\Gamma^{(1)}\,\mathbb{1}_d$,

$$\nu^* = \nu + \tfrac{1}{2}\mathbb{1}_p^T\left\{(\Sigma + M \odot M) \odot \Gamma^{(1)}\right\}\tau^{(1)} \qquad \triangleright \mathbf{E}_q\left(\boldsymbol{\sigma^{-2}} \mid \boldsymbol{y}\right)$$

  $\tau^{(1)} \leftarrow \eta^* \oslash \kappa^*$,   where $\eta^* = \eta + \frac{n}{2}\mathbb{1}_d + \frac{1}{2}\left(\Gamma^{(1)}\right)^T\mathbb{1}_p$,

$$\kappa^* = \kappa + \tfrac{1}{2}\,(y \odot y)^T\,\mathbb{1}_n - \left\{X\left(M \odot \Gamma^{(1)}\right) \odot y\right\}^T\mathbb{1}_n$$

$$+ \left(\textstyle\sum_{s=1}^{p-1} A_s \odot \sum_{j=s+1}^{p} A_j\right)^T\mathbb{1}_n$$

$$+ \tfrac{1}{2}\left\{n - 1 + \left(\sigma^{-2}\right)^{(1)}\right\}\left\{\Gamma^{(1)} \odot (\Sigma + M \odot M)\right\}^T\mathbb{1}_p,$$

$$A_s = X_s\left(M \odot \Gamma^{(1)}\right)_{s\cdot} \qquad\qquad\qquad \triangleright \mathbf{E}_q\left(\boldsymbol{\tau_t} \mid \boldsymbol{y}\right)$$

  $\Sigma \leftarrow \mathbb{1}_p\mathbb{1}_d^T \oslash B$,   where $B = \left\{n - 1 + \left(\sigma^{-2}\right)^{(1)}\right\}\mathbb{1}_p\left(\tau^{(1)}\right)^T$ $\qquad \triangleright \mathbf{Var}_q\left(\boldsymbol{\beta_{st}} \mid \boldsymbol{\gamma_{st}} = \mathbf{1}, \boldsymbol{y}\right)$

  $\log(\tau)^{(1)} \leftarrow \Psi(\eta^*) - \log(\kappa^*)$

  $\log\left(\sigma^{-2}\right)^{(1)} \leftarrow \Psi(\lambda^*) - \log(\nu^*)$

  $\log(\omega)^{(1)} \leftarrow \Psi\left(a + \Gamma^{(1)}\mathbb{1}_d\right) - \Psi\left(a + b + d\mathbb{1}_p\right)$

  $\log(1 - \omega)^{(1)} \leftarrow \Psi\left(b - \Gamma^{(1)}\mathbb{1}_d + d\mathbb{1}_p\right) - \Psi\left(a + b + d\mathbb{1}_p\right)$

  for $s = 1, \ldots, p$ **do**

    for $t = 1, \ldots, d$ **do**

      $M_{st} \leftarrow \Sigma_{st}\tau_t^{(1)}X_s^T\left(y_t - \sum_{j=1, j\neq s}^{p}\Gamma_{jt}^{(1)}M_{jt}X_j\right)$ $\qquad \triangleright \mathbf{E}_q\left(\boldsymbol{\beta_{st}} \mid \boldsymbol{\gamma_{st}} = \mathbf{1}, \boldsymbol{y}\right)$

      $\Gamma_{st}^{(1)} \leftarrow \left[1 + (\Sigma_{st})^{-1/2}\right.$

        $\left.\times \exp\left\{\log(1 - \omega_s)^{(1)} - \log(\omega_s)^{(1)} - \tfrac{1}{2}\log(\tau_t)^{(1)} - \tfrac{1}{2}\log(\sigma^{-2})^{(1)} - \tfrac{1}{2}(M_{st})^2(\Sigma_{st})^{-1}\right\}\right]^{-1}$

$$\triangleright \mathbf{E}_q\left(\boldsymbol{\gamma_{st}} \mid \boldsymbol{y}\right)$$

    end for

  end for

  $\omega^{(1)} \leftarrow a^* \oslash (a^* + b^*)$,   where $a^* = a + \Gamma^{(1)}\mathbb{1}_d$, $b^* = b - \Gamma^{(1)}\mathbb{1}_d + d\mathbb{1}_p$ $\qquad \triangleright \mathbf{E}_q\left(\boldsymbol{\omega_s^{(1)}} \mid \boldsymbol{y}\right)$

  $\mathcal{L}^{\text{old}}(q) \leftarrow \mathcal{L}(q)$, it $\leftarrow$ it $+ 1$

  Compute $\mathcal{L}(q)$ (see Appendix B.2) based on the current parameter updates

---

until $|\mathcal{L}(q) - \mathcal{L}^{\text{old}}(q)| < $ tol or it $=$ maxit

---

The symbols $\odot$ and $\oslash$ are the Hadamard operators standing for element-wise multiplication and division of two matrices of the same dimension.

B.4 Computational details. The convergence characteristics of our procedure can be described in terms of those of any deterministic iterative algorithm. In our experiments, we set the tolerance for the stopping criterion to 10^{-6}, as our empirical tests suggest that when using a smaller tolerance, the additional time required until convergence does not yield noticeably better inferences. While MCMC sampling may require thousands of iterations to converge, our algorithm usually converges in tens of iterations. As suggested by the runtime profiling provided in Appendix D.3, inference for typical genome-wide association problems with multiple outcomes is usually completed in hours. Our algorithm requires the initialization of the variational parameters, but unfortunately comes with no guarantee that it will attain a global minimum for the Kullback–Leibler divergence, $\mathrm{KL}\,(q\|p)$. This drawback can be alleviated by using several different initializations, at the price of increasing the computational effort. In practice we did not encounter situations where different starting points gave different optima. The source code can be found in the publicly available R package `locus`.

C   Details on the empirical quality assessment of the variational approximation

C.1 Marginal likelihood computation. We have

$$
\begin{aligned}
p(y) &= \int \cdots \int \mathrm{d}\omega \, \mathrm{d}\sigma^{-2} \, p(\omega) \, p\left(\sigma^{-2}\right) \\
&\qquad \times \prod_{t=1}^{d} \left\{ \sum_{\gamma_t \in \{0,1\}^p} p\left(\gamma_t \mid \omega\right) \int \cdots \int \mathrm{d}\beta_t \, \mathrm{d}\tau_t \, p\left(y_t \mid \beta_t, \tau_t\right) \, p\left(\beta_t \mid \gamma_t, \tau_t, \sigma^{-2}\right) \, p(\tau_t) \right\} \\
&= \int \cdots \int \mathrm{d}\omega \, \mathrm{d}\sigma^{-2} \left\{ \prod_{s=1}^{p} p(\omega_s) \right\} p\left(\sigma^{-2}\right) \prod_{t=1}^{d} \left\{ \sum_{\gamma_t \in \{0,1\}^p} p\left(y_t \mid \gamma_t, \sigma^{-2}\right) \prod_{s=1}^{p} p\left(\gamma_{st} \mid \omega_s\right) \right\}, \qquad (17)
\end{aligned}
$$

and one can obtain a closed form expression for $p\left(y_t \mid \gamma_t, \sigma^{-2}\right)$, after integrating out β_t and then τ_t. Indeed, proceeding similarly as in George and McCulloch (1997),

$$
\begin{aligned}
p\left(y_t \mid \tau_t, \gamma_t, \sigma^{-2}\right) &= \int (2\pi)^{-n/2} \tau_t^{n/2} \exp\left\{ -\frac{\tau_t}{2} \|y_t - X\beta_t\|^2 \right\} (2\pi)^{-q_{\gamma_t}/2} \sigma^{-q_{\gamma_t}} \tau_t^{q_{\gamma_t}/2} \exp\left\{ -\frac{\tau_t}{2}\sigma^{-2}\|\beta_t\|^2 \right\} \mathrm{d}\beta_t \\
&= \int (2\pi)^{-n/2 - q_{\gamma_t}/2} \tau_t^{n/2 + q_{\gamma_t}/2} \exp\left\{ -\frac{\tau_t}{2} \left(\beta_{\gamma_t} - \mu_{\beta_t}\right)^T V_{\gamma_t, \sigma^{-2}} \left(\beta_{\gamma_t} - \mu_{\beta_t}\right) \right\} \\
&\qquad \times \exp\left(-\frac{\tau_t}{2} S_{\gamma_t}^2 \right) \sigma^{-q_{\gamma_t}} \mathrm{d}\beta_t \,,
\end{aligned}
$$

where

$$
\begin{aligned}
q_{\gamma_t} &= \sum_{s=1}^{p} \gamma_{st}, \qquad \tilde{X}_{\gamma_t} = \begin{pmatrix} X_{\gamma_t} \\ \sigma^{-1} I_{q_{\gamma_t}} \end{pmatrix}, \qquad \tilde{y}_t = \begin{pmatrix} y_t \\ 0 \end{pmatrix}, \\
S_{\gamma_t, \sigma^{-2}}^2 &= \|\tilde{y}_t\|^2 - \tilde{y}_t^T \tilde{X}_{\gamma_t} \left(\tilde{X}_{\gamma_t}^T \tilde{X}_{\gamma_t} \right)^{-1} \tilde{X}_{\gamma_t}^T \tilde{y}_t = \|y_t\|^2 - y_t^T X_{\gamma_t} V_{\gamma_t, \sigma^{-2}}^{-1} X_{\gamma_t}^T y_t \,, \\
V_{\gamma_t, \sigma^{-2}} &= \tilde{X}_{\gamma_t}^T \tilde{X}_{\gamma_t} = X_{\gamma_t}^T X_{\gamma_t} + \sigma^{-2} I_{q_{\gamma_t}}, \qquad \mu_{\beta_t} = V_{\gamma_t, \sigma^{-2}}^{-1} \tilde{X}_{\gamma_t}^T \tilde{y}_t \,. \qquad (18)
\end{aligned}
$$

Hence, if $q_{\gamma_t} \neq 0$,

$$
p\left(y_t \mid \tau_t, \gamma_t, \sigma^{-2}\right) = (2\pi)^{-n/2} \tau_t^{n/2} \det\left(V_{\gamma_t, \sigma^{-2}}\right)^{-1/2} \exp\left(-\frac{\tau_t}{2} S_{\gamma_t}^2 \right) \sigma^{-q_{\gamma_t}} \,.
$$

Now,

$$
\begin{aligned}
p\left(y_t \mid \gamma_t, \sigma^{-2}\right) &= \int p\left(y_t \mid \tau_t, \gamma_t, \sigma^{-2}\right) p(\tau_t) \mathrm{d}\tau_t \\
&= \int (2\pi)^{-n/2} \tau_t^{n/2} \det\left(V_{\gamma_t,\sigma^{-2}}\right)^{-1/2} \exp\left(-\frac{\tau_t}{2} S_{\gamma_t}^2\right) \sigma^{-q_{\gamma_t}} \frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)} \tau_t^{\eta_t-1} \exp\left\{-\kappa_t \tau_t\right\} \mathrm{d}\tau_t \\
&= (2\pi)^{-n/2} \det\left(V_{\gamma_t,\sigma^{-2}}\right)^{-1/2} \Gamma\left(\frac{n}{2}+\eta_t\right) \frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)} \left(\kappa_t + \frac{S_{\gamma_t}^2}{2}\right)^{-n/2-\eta_t} \left(\sigma^{-2}\right)^{q_{\gamma_t}/2}.
\end{aligned}
$$

If $q_{\gamma_t} = 0$, then

$$
p\left(y_t \mid \gamma_t, \sigma^{-2}\right) = (2\pi)^{-n/2} \Gamma\left(\frac{n}{2}+\eta_t\right) \frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)} \left(\kappa_t + \frac{\|y_t\|^2}{2}\right)^{-n/2-\eta_t}.
$$

C.2 Simple Monte Carlo posterior quantities. The marginal posterior probability of inclusion for covariate X_s and response y_t can be approximated using simple Monte Carlo sums, as follows,

$$
\begin{aligned}
p(\gamma_{st} = 1 \mid y) &= \frac{p(\gamma_{st}=1, y)}{p(y)} \\
&= \frac{1}{p(y)} \int \cdots \int \mathrm{d}\omega\, \mathrm{d}\sigma^{-2} \left\{\prod_{s=1}^{p} p(\omega_s)\right\} p\left(\sigma^{-2}\right) \times \\
&\quad \left[\prod_{t'\neq t}\left\{\sum_{\gamma_{t'}\in\{0,1\}^p} p\left(y_{t'} \mid \gamma_{t'}, \sigma^{-2}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't'} \mid \omega_{s'}\right)\right\} \times \right. \\
&\quad \left. \left\{\sum_{\gamma_t\in\{0,1\}^p:\, \gamma_{st}=1} p\left(y_t \mid \gamma_t, \sigma^{-2}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}\right)\right\}\right] \\
&= \frac{1}{p(y)} \frac{1}{I} \sum_{i=1}^{I} \left[\prod_{t'\neq t}\left\{\sum_{\gamma_{t'}\in\{0,1\}^p} p\left(y_{t'} \mid \gamma_{t'}, (\sigma^{-2})^{(i)}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't'} \mid \omega_{s'}^{(i)}\right)\right\} \times \right. \\
&\quad \left. \left\{\sum_{\gamma_t\in\{0,1\}^p:\, \gamma_{st}=1} p\left(y_t \mid \gamma_t, (\sigma^{-2})^{(i)}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}^{(i)}\right)\right\}\right],
\end{aligned}
$$

where the samples are generated independently from

$$
\left(\sigma^{-2}\right)^{(i)} \sim \mathrm{Gamma}(\lambda, \nu), \qquad \omega_s^{(i)} \sim \mathrm{Beta}(a_s, b_s), \quad s = 1, \ldots, p, \qquad i = 1, \ldots, I. \tag{19}
$$

Similarly, we approximate the posterior mean for ω_s as

$$
\begin{aligned}
\mathrm{E}(\omega_s \mid y) &= \int \omega_s\, p(\omega_s \mid y) \mathrm{d}\omega_s = \frac{1}{p(y)} \int \omega_s\, p(\omega_s, y) \mathrm{d}\omega_s \\
&= \frac{1}{p(y)} \int \cdots \int \mathrm{d}\omega\, \mathrm{d}\sigma^{-2}\, \omega_s \left\{\prod_{s'=1}^{p} p(\omega_{s'}')\right\} p\left(\sigma^{-2}\right) \times \\
&\quad \prod_{t=1}^{d}\left\{\sum_{\gamma_t\in\{0,1\}^p} p\left(y_t \mid \gamma_t, \sigma^{-2}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}\right)\right\} \\
&= \frac{1}{p(y)} \frac{1}{I} \sum_{i=1}^{I} \omega_s^{(i)} \prod_{t=1}^{d}\left\{\sum_{\gamma_t\in\{0,1\}^p} p\left(y_t \mid \gamma_t, (\sigma^{-2})^{(i)}\right) \prod_{s'=1}^{p} p\left(\gamma_{s't} \mid \omega_{s'}^{(i)}\right)\right\}.
\end{aligned}
$$

Figures 7, 8, 9, and 10 display and compare diverse posterior quantities obtained by variational, MCMC or simple Monte Carlo approximations.
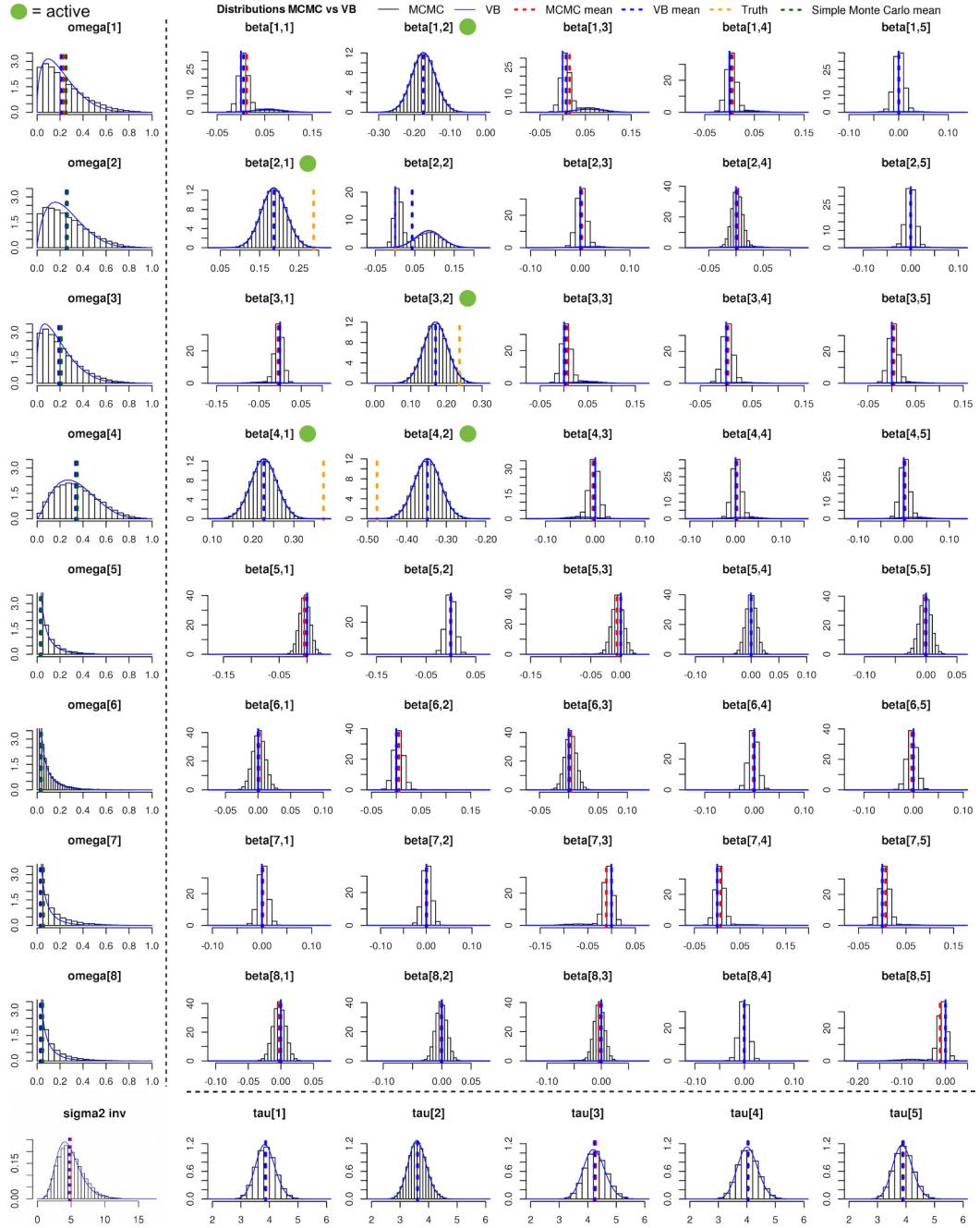

Figure 7: MCMC histograms and variational Bayes (VB, blue) posterior densities for parameters ω (far left panels), β (central panels), σ^{-2} (bottom far left panel) and τ (bottom panels). The MCMC means (dashed red) and variational means (dashed blue) are displayed, along with the simulated values for the β plots (dashed orange) and the simple Monte Carlo approximation of the posterior mean of ω (dashed green). Most of the dashed vertical lines overlap. The problem has of $p = 8$ independent covariates and $d = 5$ responses for $n = 250$ samples. The five green dots indicate the simulated nonzero associations; each explains on average 13.5% of response variance. We use the software OpenBUGS (Spiegelhalter et al., 2007) and the R package coda (Plummer et al., 2006) for the MCMC inference and convergence diagnostics.
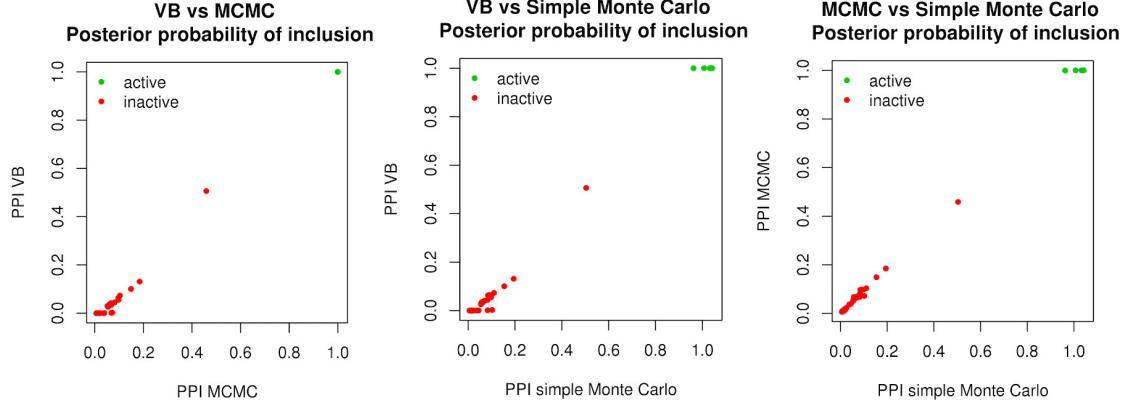
Figure 8: Comparison of marginal posterior probabilities of inclusion (PPI) for variational Bayes (VB), MCMC, and simple Monte Carlo approximations. The posterior probabilities of inclusion corresponding to the true signals all overlap, the smallest value being 0.99991 for variational and 0.99926 for MCMC inferences. A better simple Monte Carlo approximation might be obtained by increasing the number of draws (here $I = 2 \times 10^5$).
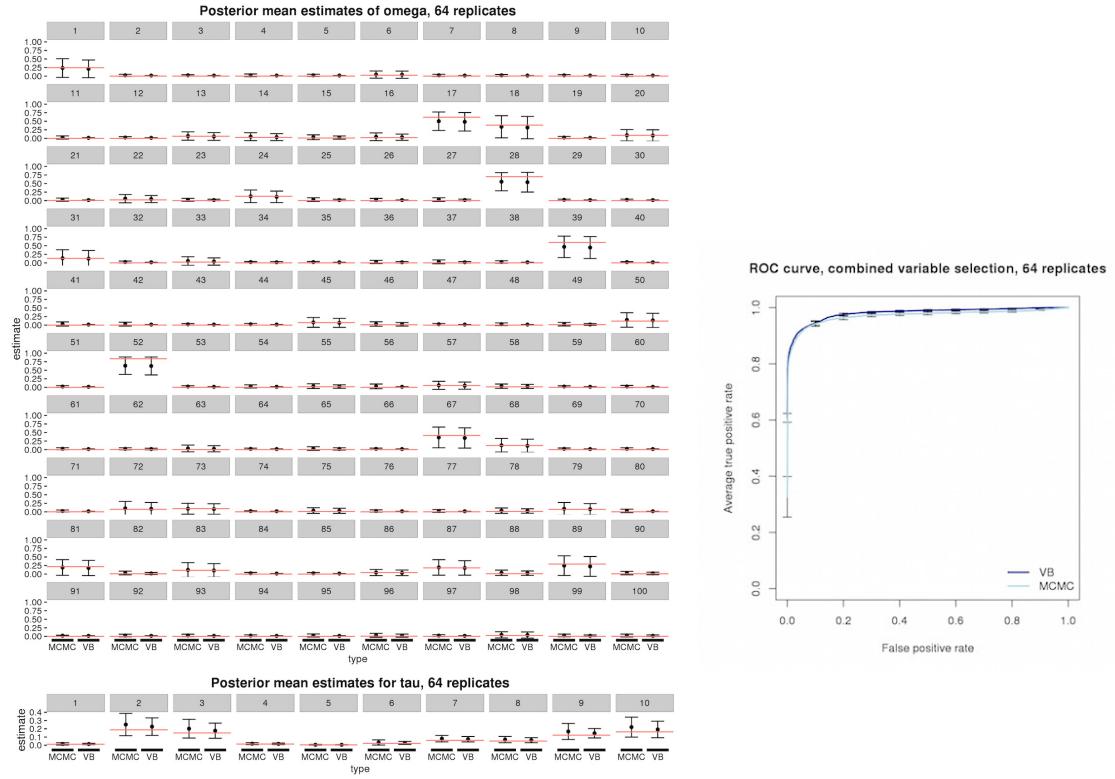

Figure 9: Comparison of variational and MCMC inferences on data generated from the model with $p = 100$ covariates, $d = 10$ responses and $n = 50$ samples. Left: posterior means of ω (central panels) and τ (bottom) with confidence intervals from 64 replications. The confidence intervals are similar, and most of them contain the simulated value of the parameter (red horizontal lines). Right: average receiver operating characteristic curves for combined variable selection based on the full marginal posterior probabilities matrix, i.e., $\{\gamma_{st}^{(1)}\}$ for variational inference and obtained by dividing the matrix of counts for $\{\gamma_{st}\}$ by the chain length minus the burn-in length for MCMC inference.
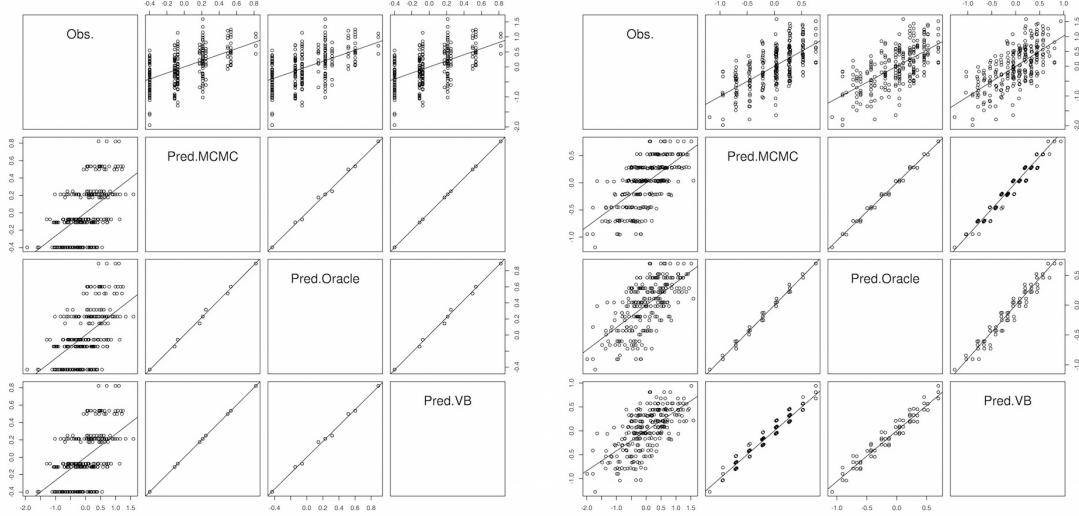
Figure 10: Observed values y_t and estimated posterior means of $X\beta_t$ obtained by variational and MCMC approximations, and an oracle, for $t = 1$ (left panels, two active covariates) and $t = 2$ (right panels, three active covariates). Simulation described in Section 4.2, $p = 8$ covariates, $d = 5$ responses and $n = 250$ samples.

D   Details on the simulation studies

D.1 Data generation design. We provide here some implementation details on the data-generation settings described in Section 5.1. The same general design is used for all the numerical experiments; it is tailored to genetic association studies with multiple outcomes. The dependence structure of simulated SNPs and molecular outcomes is by blocks. As we assume Hardy–Weinberg equilibrium for the SNPs, their marginal distribution is a binomial $B(2, m)$ with probability m equal to their respective minor allele frequencies, so we model dependence within each block using realisations from a multivariate Gaussian latent variable whose correlation matrix describes a desired dependence structure (autocorrelation with prescribed correlation coefficient or correlation structure corresponding to that of real SNPs at disposal). SNPs are then obtained using a quantile thresholding rule involving their preselected minor allele frequencies; this approach is along the lines of copula-based dependence modelling (without having to resort to copulas). When simulating SNPs that emulate real SNPs data by approximating their minor allele frequencies and correlation structure, the empirical covariance of real SNP blocks may not be positive definite, in which case we approximate it by the closest positive definite covariance matrix (in Frobenius norm) using the algorithm of Higham (2002) implemented in the R package Matrix (Bates and Maechler, 2015). Outcomes are associated with SNPs under an additive dose-effect scheme (uniform and linear increase in risk for each copy of the minor allele) and the proportions of outcome variance explained per active SNP are drawn from a $\text{Beta}(\alpha, \beta)$ distribution, with shape parameters $\alpha = 2$ and $\beta = 5$ chosen to give more weight to smaller effect sizes. These proportions are rescaled to match a preselected average proportion of outcome variance explained per active SNP (assuming that genetic and external environmental effects are uncorrelated). The magnitude of SNP effects derives from these choices and the sign of these effects is switched with probability 0.5. Such a

design implies an inverse relationship between minor allele frequencies and effect sizes, which is expected to occur since selection against SNPs with large penetrance is stronger (see, e.g., Park et al., 2011). Additional information regarding the generation of a pleiotropic association pattern is given in the paper, as it may vary with the test case considered.

D.2 Competing predictor selection methods. The performance of our approach in terms of predictor selection is compared with the following regression procedures:

- univariate ordinary least squares: each response y_t is regressed on each covariate X_s and the statistics $\max_{t=1,\ldots,d} t_{st}$, where t_{st} is the t-statistic for the significance of β_{st}, are gathered and ranked;

- elastic net regression for multivariate Gaussian responses ($\alpha = 0.5$) with 10-fold cross-validation for choosing the tuning parameter λ (glmnet, Friedman et al., 2009). The estimates $|\beta_s|$ ($s = 1,\ldots,p$), where β_s is the regression coefficient estimate for X_s and common to all responses, are gathered and ranked;

- univariate Bayesian regressions, lmBF (Morey and Rouder, 2015): each response y_t is regressed on each covariate X_s with all computations made analytically. The (average) Bayes factors, $\sum_{t=1}^d \mathrm{BF}_{st}/d$ ($s = 1,\ldots,p$), are gathered and ranked;

- d Bayesian multiple regressions, BAS (Clyde, 2016), one for each response, using MCMC inference. A g-prior is used for the regression coefficients. The (average) Bayes factors, $\sum_{t=1}^d \mathrm{BF}_{st}/d$ ($s = 1,\ldots,p$), are gathered and ranked; and

- d Bayesian multiple regressions, varbvs (Carbonetto and Stephens, 2012), one for each response, using variational inference. The posterior probabilities of inclusion are summed across responses and ranked.

D.3 Runtime profiling. We report a runtime profiling of the different methods for a range of problem sizes ($n \times p \times d$). Figure 11 displays the run times in minutes, averaged over 24 replications. We do not aim to provide precise and exhaustive comparisons, since the methods all depend on parallelism and convergence characteristics that are not directly comparable. All methods were run serially, in an attempt to treat them on an equal footing. This choice could be challenged, as some approaches are more parallelizable than others, but the number of cores used for the latter represents an additional setting that would come into play with a potentially large impact on the measures. The number of chains for MCMC inferences also matters: HESS (Richardson et al., 2010) is run with three chains (following its authors' choice made in their simulations); the other MCMC inferences are based on a single chain. Finally, the runtime may also greatly vary depending on the chosen chain length: the latter is adaptively selected by the Bayesian multiple regression method BAS (Clyde, 2016), and, based on preliminary convergence diagnostics, it is set to $50,000$ samples for HESS, iBMQ (Scott-Boyer et al., 2012) and the MCMC inference on our model using OpenBUGS (Spiegelhalter et al., 2007). In practice, the number of samples needed until convergence may increase greatly with the problem size, a fact that was not accounted for in this profiling. Hence, by adequately increasing the chain lengths when considering larger dimensionalities, we expect the curves of Figure 11 corresponding to the MCMC approaches
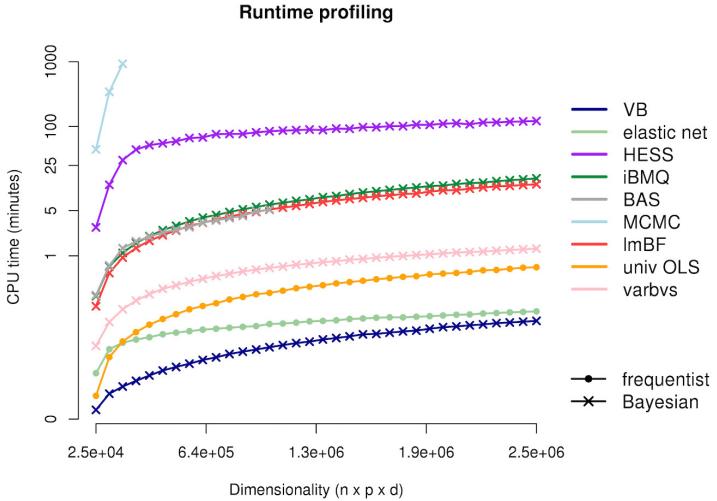
Runtime profiling

Figure 11: Serial runtime profiling for all methods discussed, on an Intel Xeon CPU at 2.60 GHz with 64 GB RAM.

to deviate more widely from that of our method. With these serial settings, our approach is faster than all Bayesian and frequentist methods. MCMC inference for our model is the slowest of all tested methods, which underlines the intractability of MCMC sampling for large problems. Inference for HESS is faster, but still more than 650 times slower than our variational approach. Our method is also about 10 times faster than d applications (one for each outcome) of the varbvs method (Carbonetto and Stephens, 2012) but the runtime of the varbvs procedure can be reduced using multiple cores.

E   Details on the real data problem

E.1 Clinical study and preprocessing of the mQTL data analysis. Diogenes (Larsen et al., 2010) is a large clinical dietary intervention study, which enrolled roughly $1,000$ overweight subjects from 8 European centers. The subjects were assigned to a 8-week weight-loss program followed by 6 months of supervised ad-libitum diet, meant as a weight-maintenance phase. For the 6-month period, individuals were randomized into five intervention groups, whose diets differed in their macronutrient and glycemic load content. The clinical trial had the broad objective of elucidating whether certain macronutrient compositions are more successful in weight maintenance than others. It gathered data on genetic variants, gene, protein and metabolite expression. In addition, more than $7,000$ clinical, anthropometric and behavioural variables were made available for each individual. When meaningful to do so, the data were collected at three time points, reflecting different stages of the dietary treatment.

In Section 5.3, we performed an mQTL analysis which involves tag SNPs from Illumina Human-Core chips (about 300k SNPs) and metabolites from plasma obtained by liquid chromatography-mass spectrometry (LC-MS). We excluded SNPs having missing values, minor allele frequency below 5%, call rate below 95%, or violating the Hardy–Weinberg equilibrium. Moreover, subjects having gender discrepancy (e.g., subject recorded as male but being homozygous for each X chromosome marker), abnormal autosomal heterozygosity or whose genomes were too close to each

other (IBS> 95%) were also excluded. Additional subjects were removed after applying the Tukey method for outlier detection (Tukey, 1949), based on the metabolomic data. Metabolite expression levels were log2-transformed, and had no missing values. After these quality checks, the data consisted of $p = 215'907$ SNPs and $d = 125$ metabolite expression levels, for $n = 317$ individuals.

E.2 Permutation-based Bayesian false discovery rate estimation. We provide details on the false discovery rate estimation procedure applied in Section 5.3 to compare our method with the varbvs method of Carbonetto and Stephens (2012) on the real data. We use the two-group mixture approach proposed by Efron (2008) in the context of microarray data analysis: we simultaneously consider N null hypotheses and their corresponding test statistics, which we assume to follow a mixture distribution

$$F = \pi_0 F_0 + (1 - \pi_0) F_1 \,,$$

where π_0 is the prior probability for a null case, and F_0 and F_1 are the null and non-null cumulative distribution functions. The "Bayesian false discovery rate" for some threshold τ is the posterior probability that a rejected null hypothesis (i.e., test statistic exceeding τ) is a false positive,

$$\mathrm{FDR}(\tau) = \frac{\pi_0 \bar{F}_0(\tau)}{\bar{F}(\tau)} \,, \tag{20}$$

where $\bar{F} = 1 - F$ and $\bar{F}_0 = 1 - F_0$. We derive estimates of (20) based on the posterior probabilities of inclusion quantifying the associations between each covariate-response pair. Specifically, we obtain an empirical null distribution by running our algorithm (with the same hyperparameters as those used for the actual inference) on B datasets with permuted outcome sample labels and compute, for a grid of thresholds $0 < \tau_1 < \ldots < \tau_K < 1$,

$$\widehat{\mathrm{FDR}}(\tau_k) = \frac{\mathrm{median}_{b=1,\ldots,B} \#\{\mathrm{PPI}_{st}^{(b)} > \tau_k; \ s = 1,\ldots,p; \ t = 1,\ldots,d\}}{\#\{\mathrm{PPI}_{st} > \tau_k; \ s = 1,\ldots,p; \ t = 1,\ldots,d\}}, \qquad k = 1,\ldots,K, \tag{21}$$

where we conservatively set π_0 to 1 in (20). As the posterior probabilities of inclusion obtained by applying varbvs d times (one multiple regression for each outcome) are not identically calibrated across all d estimations, we use adaptive thresholds on the columns of the varbvs posterior probabilities of inclusion matrix,

$$\tau_k(y_t) = \frac{\mathrm{median}_s \, (\mathrm{PPI}_{st})}{\mathrm{median}_{s,t'} \, (\mathrm{PPI}_{st'})} \, \tau_k \qquad t = 1,\ldots,d, \quad k = 1,\ldots,K.$$

To find thresholds $\hat{\tau}$ corresponding to preselected false discovery rates, we fit a cubic spline to the estimates (21) previously obtained for $\tau_1, \ldots, \tau_K$.

E.3 Biological evidence for the mQTL analysis findings. We support the findings obtained by our method for the mQTL data analysis with external association results from the following online databases:

- GWAS catalog (Welter et al., 2014);

- UCSC genome browser (Karolchik et al., 2003);

- GTEx (Lonsdale et al., 2013); and

- GeneCards (Rebhan et al., 1998).

Replication mQTL problem, ROC curves
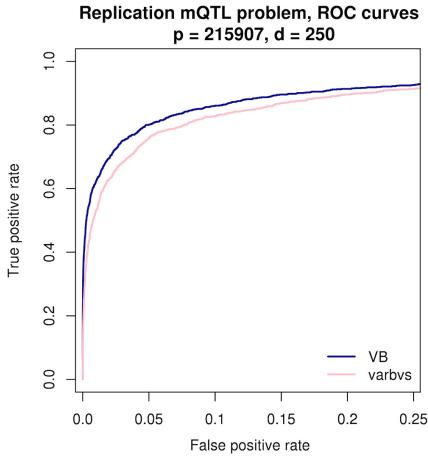**p = 215907, d = 250**

Figure 12: Receiver operating characteristic curves based on posterior probabilities of inclusion inferred by VB and by varbvs for simulated mQTL data.

We find direct or indirect links with metabolic activities for 12 of the 25 SNPs declared as "active' by our method: SNPs $rs3820711$, $rs4316911$, $rs4909818$, $rs4744227$, $rs174535$, $rs680379$, $rs8012466$, $rs4906771$, $rs573922$, $rs3903703$, $rs8114788$ and $rs6001093$ have been identified as associated with BMI, diverse diabetic or obesity diseases, fatty acid, sphingolipid or phospholipid levels, either from direct genome-wide association analyses or through protein coding genes for which they were reported as eQTLs.

E.4 Replication of the mQTL data analysis. In this appendix, we replicate the real data analysis of Section 5.3 on a simulated dataset (with twice as many outcomes and slightly more observations), which can be found on Figshare (`https://dx.doi.org/10.6084/m9.figshare.4509755.v1`). The analysis can be reproduced using the code available on GitHub (`https://github.com/hruffieux/mQTL_analysis_example`), provided that a machine with adequate RAM memory (about 400G for this large problem) is used.

To best mimic the real mQTL data used in Section 5.3, we simulate $p = 215,907$ SNPs based on the sample minor allele frequencies of the real tag SNPs and we reproduce their dependence structure by blocks of $1,000$ consecutive SNPs according to the discussion of Appendix D.1. We also simulate $d = 250$ normally distributed outcomes with equicorrelation by blocks using correlation coefficient $\rho \in \{0.5, 0.6, 0.7\}$; this block structure is similar to that of the real metabolite data. To induce a realistic pleiotropic pattern, we simulate associations between 750 SNPs and 175 metabolites (randomly chosen) taking the block-wise dependence structure of the latter into account: the probability that a given active SNP is associated with a given active outcome varies across blocks, so correlated metabolites are either all highly likely or all less likely to be associated with the SNP. The average proportion of metabolic variance explained with each association is 2.5% (but more associations explain less than this, as discussed in Appendix D.1). We generate $n = 350$ observations.

We apply the permutation analysis described in Section 5.3 to varbvs and our method and again obtain a more powerful selection for our method compared to varbvs at estimated FDR of 20% and 25% (Table 5 and Figure 13). Because simulated data are used, we can further support the overall superiority of our method with ROC curves, see Figure 12.

Table 5: Replication of the mQTL data analysis. Number of true positives (TP) detected by VB and by varbvs, and number of TP in common at selected permutation-based false discovery rates.

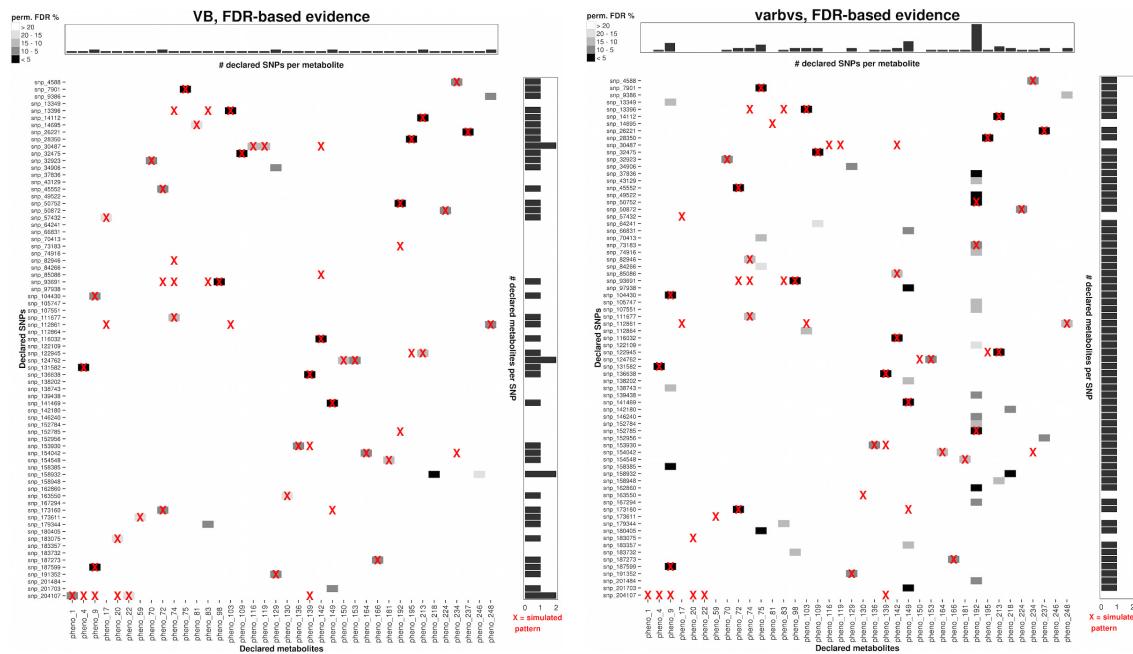| Permutation-based FDR (%) | # TP: | | |
	VB	varbvs	VB ∩ varbvs
5	13	18	13
10	26	26	23
15	32	32	28
20	38	32	28
25	47	33	29

Figure 13: Comparison of the associations declared by VB (left) and varbvs (right) at FDR of 20% estimated using $B = 200$ permutations for simulated mQTL data; see also Table 5. The simulated association pattern is overlaid (red crosses). The large number of false positives on the right plot indicates that the permutation-based FDR estimates are somewhat underestimated for varbvs; an improved estimation based on more permutations would further emphasize the improvement of our method over varbvs.
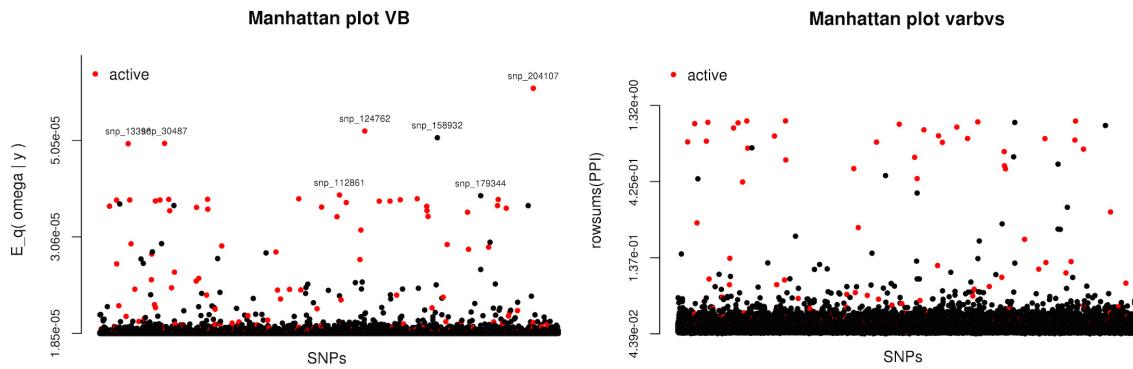

Figure 14: Manhattan plots of SNP association and evidence of pleiotropy obtained with VB (left) and varbvs (right) for simulated mQTL data.

36