Biostatistics (2017) **18**, 4, *pp.* 618–636
doi:10.1093/biostatistics/kxx007
Advance Access publication on March 16, 2017

Efficient inference for genetic association studies with multiple outcomes

HELENE RUFFIEUX*

Nestlé Institute of Health Sciences SA, EPFL Innovation Park, 1015 Lausanne, Switzerland
Ecole Polytechnique Fédérale de Lausanne, EPFL SB MATH STAT, Station 8,
1015 Lausanne, Switzerland
helene.ruffieux@rd.nestle.com

ANTHONY C. DAVISON

Ecole Polytechnique Fédérale de Lausanne, EPFL SB MATH STAT, Station 8,
1015 Lausanne, Switzerland
anthony.davison@epfl.ch

JORG HAGER, IRINA IRINCHEEVA

Nestlé Institute of Health Sciences SA, EPFL Innovation Park, 1015 Lausanne, Switzerland
jorg.hager@rd.nestle.com, irina.irincheeva@rd.nestle.com

Summary

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 modeling are either too simplistic to be powerful or are impracticable for computational reasons. Inspired by Richardson *and others* (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.

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

*To whom correspondence should be addressed.

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 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 *and others*, 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 *and others*, 2011; Solovieff *and others*, 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 modeling of multiple outcomes in genetic association problems have recently been made. Flutre *and others* (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 *and others* (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 *and others*, 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 *and others* (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 modeling 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 modeling possibilities in which biological beliefs may be naturally incorporated. The hierarchical regression approach of Richardson *and others* (2010), hierarchical evolutionary stochastic search (HESS), 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 *and others* (2012) propose methods called BAYES and integrated Bayesian hierarchical model for eQTL mapping (iBMQ) based on models similar to that of HESS, but a major drawback of all three approaches is the lack of scalability of their MCMC 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 article, 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 MCMC approaches. We describe a variational inference procedure for a model similar to that of Richardson *and others* (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 article 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 MCMC 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 *and others*, 2010) and iBMQ (Scott-Boyer *and others*, 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 article 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 *and others*, 2010). Suppose that

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

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 modeling 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 (PPI), i.e., the posterior means of $\{\gamma_{st}\}$.

Our model differs from that of Richardson *and others* (2010) in two respects. One concerns the treatment of the regression coefficient parameters β_{st}: we use independent priors, whereas Richardson *and others* 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 *and others* (2010) opt for a quite complex specification, in which

$$\omega_{st} = \rho_s \omega_t, \qquad \omega_t \sim \text{Beta}(a_t, b_t), \qquad \rho_s \sim \text{Gamma}(c_s, d_s), \qquad 0 \leq \omega_{st} \leq 1. \tag{2.1}$$

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 *and others* (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,

$$\omega_{st} \mid a_s, b_s, \pi_s \sim \pi_s \delta_0 + (1 - \pi_s)\text{Beta}(a_s, b_s)\,, \qquad \pi_s \sim \text{Beta}(a_0, b_0)\,, \tag{2.2}$$

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 (2.1). While such a formulation was mentioned by Richardson *and others* (2010) and by Scott-Boyer *and others* (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), λ, $\nu \in \mathbb{R}_+$ (for σ^{-2}'s Gamma prior) and η, $\kappa \in \mathbb{R}_+^d$ (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 favor 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 article, 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 favoring 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

$$a_s \equiv 1\,, \qquad b_s \equiv d(p - p^*)/p^*\,, \qquad 0 < p^* < p\,, \tag{2.3}$$

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 favored 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.,

$$\text{POR}(q_s - 1 : q_s) = \frac{p\left(\sum_{t=1}^d \gamma_{st} = q_s - 1\right)}{p\left(\sum_{t=1}^d \gamma_{st} = q_s\right)} = \frac{b_s + d - q_s}{a_s + q_s - 1}\,, \qquad q_s = 1, \ldots, d\,. \tag{2.4}$$

Clearly, penalties arise and increase with the total number of responses in the model, d. Figure 1 displays (2.4) for $q_s = 1, \ldots, 5$ as a function of p and indicates that, when a_s and b_s are specified as in (2.3), 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 (2.3) to an "uncorrected" regime with $a_s \equiv 1$, $b_s \equiv 2d - 1$, so that the prior mean number of responses associated

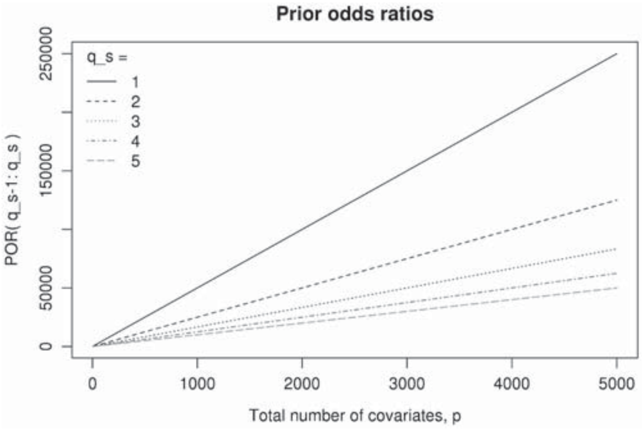

Fig. 1. Prior odds ratios, $\mathrm{POR}(q_s - 1 : q_s)$, for $q_s = 1, \ldots, 5$, a_s and b_s as in (2.3), $d = 100$, $p^* = 2$, and for a total number of covariates ranging from $p = 5$ to 5000; see Scott and Berger (2010) for a similar visualization of prior odds ratios in a single response context.

Table 1. *Multiplicity adjustment at covariate level. The mean number 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	1000	2500
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)

with X_s is 0.5, i.e., $\mathrm{E}(\omega_s) \equiv (2d)^{-1}$. The number of false positives (FP), 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 (2.3), giving a clear multiplicity adjustment. Other experiments confirmed strong sparsity control; the reported findings on real data should therefore be plausible when (2.3) is used.

3. Variational inference

Section 2 described some differences between our model and those of Jia and Xu (2007, BAYES), Richardson *and others* (2010, HESS), and Scott and Berger (2010, iBMQ), but a more fundamental distinction concerns the inference procedure. The three earlier methods rely on 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 *and others*, 2002; Attias, 2000) to construct such a class, i.e., we assume that $q(\theta)$ factorizes over some partition of θ, $\{\theta_j\}_{j=1,\ldots,J}$,

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

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

$$q\left(\beta, \gamma, \tau, \sigma^{-2}, \omega\right) = \left\{\prod_{s=1}^{p}\prod_{t=1}^{d} q(\beta_{st}, \gamma_{st})\right\}\left\{\prod_{s=1}^{p} q(\omega_s)\right\}\left\{\prod_{t=1}^{d} q(\tau_t)\right\} q\left(\sigma^{-2}\right), \tag{3.1}$$

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

$$\log p(y) = \mathcal{L}(q) + \mathrm{KL}\,(q \,\|\, p)\;, \tag{3.2}$$

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

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{k=1}^{J} q_k(\theta_k)\left\{\log p(y,\theta) - \sum_{k=1}^{J} \log q_k(\theta_k)\right\} \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_J \\ &= \int q_j(\theta_j)\left\{\int \log p(y,\theta) \prod_{k \neq j} q_k(\theta_k)\mathrm{d}\theta_k - \log q_j(\theta_j)\right\} \mathrm{d}\theta_j + \mathrm{cst} \\ &= \int q_j(\theta_j) \log\left\{\frac{p_{-j}(\theta_j; y)}{q_j(\theta_j)}\right\} \mathrm{d}\theta_j + \mathrm{cst}, \qquad j = 1, \ldots, J\,,\end{aligned} \tag{3.3}$$

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 (3.3) 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

$$\log q_j(\theta_j) = \mathrm{E}_{-j}\{\log p(y,\theta)\} + \mathrm{cst}, \qquad j = 1, \ldots, J\,. \tag{3.4}$$

The relations (3.4) 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 (3.1) 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

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

where the *variational parameters* $\mu_{\beta,st}, \sigma^2_{\beta,st}, \gamma^{(1)}_{st}$ 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, Sections 3.1.5, 3.2.4, 3.2.5). The algorithm and its derivation are given in Section 2 of the supplementary material available at *Biostatistics* online.

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 KL $(q \parallel p)$. Because of relation (3.2), 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

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

where

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

see Section 3 of the supplementary material available at *Biostatistics* online for details. As no closed form is available for the remaining integrals, we use

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

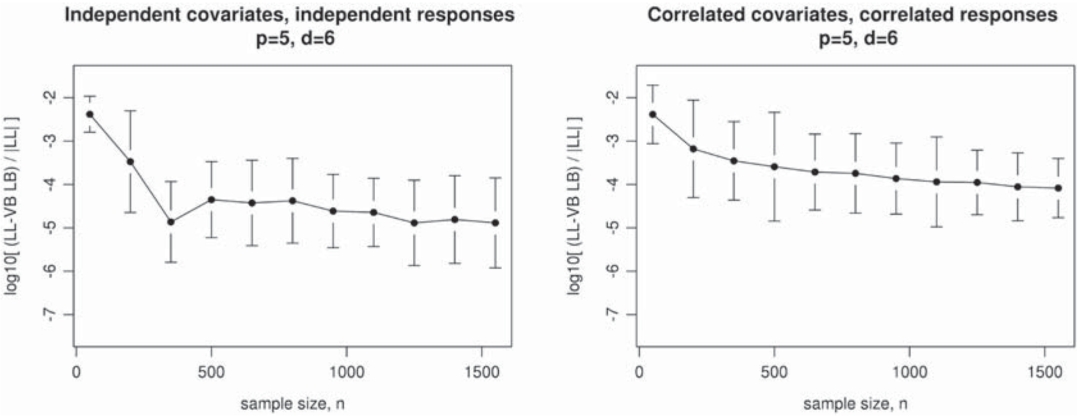

Fig. 2. Log_{10} relative difference between the marginal log-likelihood and the variational lower bound. Left: independent covariates and responses. Right: correlated covariates and responses, $\rho = 0.75$. Problems with $p = 5$ covariates, of which $p_0 = 3$ randomly selected as "active" (associated 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.

where we independently generate

$$\left(\sigma^{-2}\right)^{(i)} \sim \text{Gamma}(\lambda, \nu)\,, \qquad \omega_s^{(i)} \sim \text{Beta}(a_s, b_s)\,, \qquad s = 1, \ldots, p\,, \qquad i = 1, \ldots, I\,. \tag{4.1}$$

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 of inclusion of a

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*

			Active			Inactive
10×	$\beta_{1,2}$	$\beta_{2,1}$	$\beta_{3,2}$	$\beta_{4,1}$	$\beta_{4,2}$	β_{rest} (average)
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)
			Active			Inactive
	ω_1	ω_2	ω_3	ω_4		ω_{rest} (average)
True proportion of active responses	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

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 (4.1), 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 1 of the supplementary material available at *Biostatistics* online, 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 PPI 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 2 of the supplementary material available at *Biostatistics* online. 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 Section 3 of the supplementary material available at *Biostatistics* online. 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 modeling 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 Section 4.2 of the supplementary material available at *Biostatistics* online. 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 favor 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 Section 4.1 of the supplementary material available at *Biostatistics* online.

We perform 48 replications for each of three simulation configurations. The first configuration has moderate numbers of covariates ($p = 5000$) 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, number 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

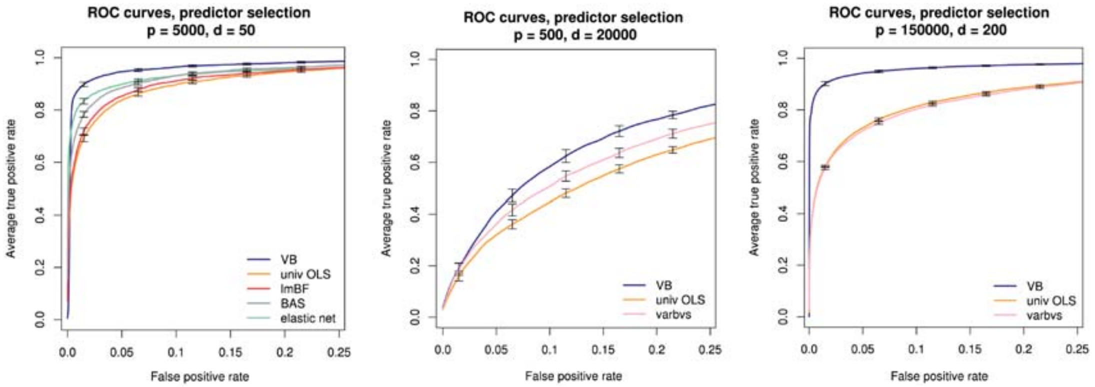

Fig. 3. Truncated average ROC curves with 95% confidence intervals for predictor selection obtained from 48 replications. The competing methods are used in three studies of different sizes, based on their computational tractability. Left: $p = 5000$ covariates spatially autocorrelated with correlation 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 $p_{\text{add}} = 0.15$, average outcome variance percentage explained by the active covariates $p_{\text{ve}} = 30.0\%$. Middle: $p = 500$ independent covariates, $d = 20\,000$ outcomes equicorrelated by blocks of size 10 with $\rho_X \in \{0.5, \dots, 0.8\}$, $p_0 = 300$, $d_0 = 12\,500$, $n = 300$, $p_{\text{add}} = 0.01$, $p_{\text{ve}} = 55.8\%$. Right: $p = 150\,000$ covariates autocorrelated by blocks of size 100 with $\rho_X \in \{0.5, \dots, 0.9\}$, $d = 200$ outcomes with same correlation structure than real protein expression levels (Diogenes study, Larsen *and others*, 2010, see Section 5.1 of the supplementary material available at *Biostatistics* online), $p_0 = 500$, $d_0 = 150$, $n = 200$, $p_{\text{add}} = 0.05$, $p_{\text{ve}} = 62.6\%$. The univariate ordinary least squares and varbvs curves overlap.

mean-field approximation, as suggested by the results of Section 4.1. The marginal ordinary least squares and marginal lmBF regressions appear to miss many associations because of their univariate modeling 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 HESS (Richardson *and others*, 2010) and iBMQ (Scott-Boyer *and others*, 2012) 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 0.05, for $n = 250$ samples. On average, the active covariates account for 22% of the variance of an outcome

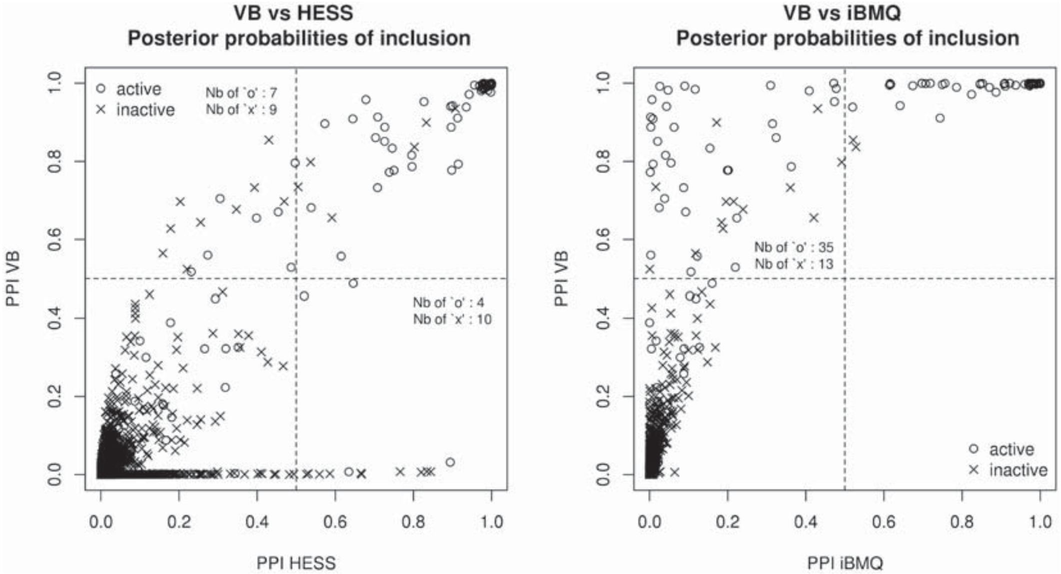

Fig. 4. Marginal PPI obtained by our approach, and those of HESS (left) and those of iBMQ (right), for a problem with $p = 250$ covariates of which $p_0 = 50$ are active, with $d = 100$ outcomes, of which $d_0 = 50$ are active, and $n = 250$ samples.

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 s with our method, around 21 min with iBMQ and 4 h 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 PPI 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 (FP) and few true positives (TP). The comparison with iBMQ is more contrasted, as the values of its PPI 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 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 PPI from the first replicate. Visual comparison of the TP 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 $\text{pr}(\rho_s > 1 \mid y)$ $(s = 1, \ldots, p)$, described

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×	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)

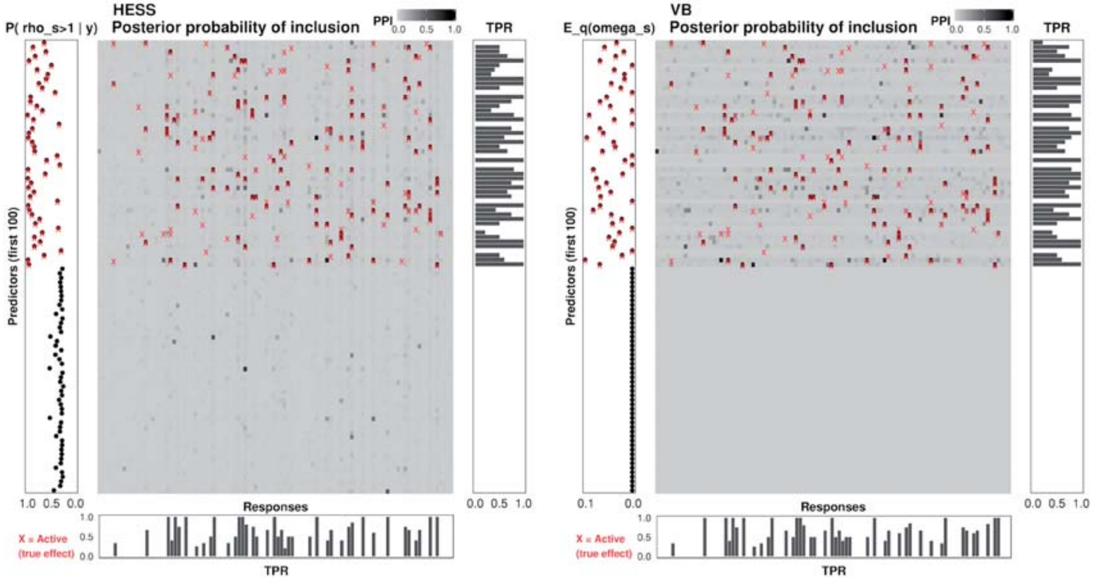

Fig. 5. Posterior quantities for detection of associations with HESS (left) and with our approach (right), for a simulated dataset with $p = 250$ independent covariates (here only the first 100 are displayed), of which $p_0 = 50$ are active and placed first, with $d = 100$ responses (50 active) and $n = 250$ individuals. Marginal PPI (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 $\text{pr}(\rho_s > 1 \mid y)$ for HESS and posterior mean $\text{E}_q(\omega_s \mid y)$ for our approach (left panel). The simulated associations are shown by crosses.

by Richardson *and others* (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 *and others*, 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

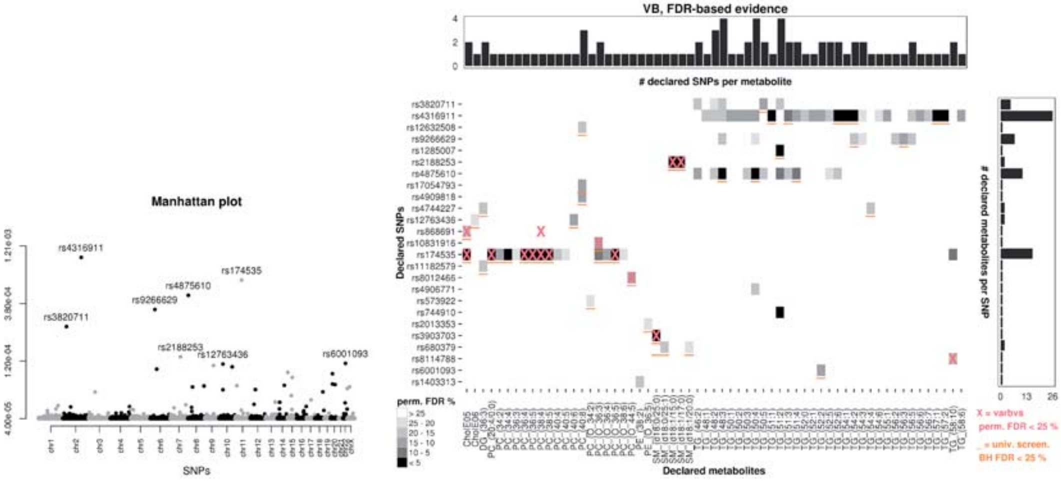

Fig. 6. SNPs and pairwise associations declared by our approach for the Diogenes study. Left: Manhattan plot for SNP association and evidence of pleiotropy. Right: PPI for SNP-metabolite associations declared at estimated FDR of 25% and overlap with the associations declared by the varbvs method at same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

trait locus (mQTL) analysis; in this context, the metabolites may be viewed as proxies for the clinical condition of interest, weight maintenance. We also use this illustration on real data to further highlight the benefits of modeling 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), phosphatidylcholines (PC), phosphatidylethanolamines (PE), sphingomyelins (SM), di- (DG) and triglycerides (TG). Section 5.1 of the supplementary material available at *Biostatistics* online 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 h on an Intel Xeon CPU at 2.60 GHz with 512 GB RAM. The posterior means $\mathrm{E}_q(\omega_s \mid y)$ suggest 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 PPI, and use an empirical null distribution based on $B = 400$ permutations to compute the estimate

$$\widehat{\mathrm{FDR}}(\tau) = \frac{\mathrm{median}_{b=1,\ldots,B}\#\{\mathrm{PPI}_{st}^{(b)} > \tau\}}{\#\{\mathrm{PPI}_{st} > \tau\}}, \qquad 0 < \tau < 1, \tag{5.1}$$

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 (univ.) 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.)

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, *rs*4316911, shows many associations with triglyceride levels, and turns out to be located less than 150*kb* from the protein coding gene ITGA6 known to be linked to diabetic kidney disease (Iyengar *and others*, 2015). The second most prominent pleiotropic SNP, *rs*174535, 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 *and others*, 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 *rs*3903703 too has been identified as associated with very long-chain fatty acid levels (Lemaitre *and others*, 2015). This seems to agree with our findings, in which *rs*3903703 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 Section 5 of the supplementary material available at *Biostatistics* online. 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 *and others*, 2010). Convergence control is automatic, whereas convergence assessment 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 modeling 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 *and others*, 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 *and others* (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 article are implemented in the publicly available R package `locus`.

SUPPLEMENTARY MATERIAL

Supplementary material available at http://biostatistics.oxfordjournals.org contains technical appendices, additional simulations, runtime profiling and details on the real data example.

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and the two referees for their very helpful comments that greatly improved the article. We thank Jérôme Carayol, Loris Michel and Armand Valsesia for their valuable comments. We also thank James Holzwarth, Bruce O'Neel and Jaroslaw Szymczak for giving us access to computing resources. *Conflict of Interest:* None declared.

FUNDING

The Diogenes trial was supported by the European Commission (FP6-2005-513946) and Nestlé Institute of Health Sciences.

REFERENCES

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

BARBIERI, M. M. AND BERGER, J. O. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 870–897.

BOYD, S. AND VANDENBERGHE, L. (2004). *Convex Optimization*. New York: Cambridge University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Mozaffarian, D., Kabagambe, E. K., Johnson, C. O., Lemaitre, R. N., Manichaikul, A., Sun, Q., Foy, M., Wang, L., Wiener, H. and Irvin, M. R. (2015). 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.

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

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

Richardson, S., Bottolo, L. and Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high-dimensional data. In: Bernardo, J. M. *and others* (editors), *Bayesian Statistics*, Volume 9. New York: Oxford University Press, pp. 539–569.

SCOTT, J. G. AND BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics* **38**, 2587–2619.

SCOTT-BOYER, M. P., IMHOLTE, G. C., TAYEB, A., LABBE, A., DESCHEPPER, C. F. AND GOTTARDO, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology* **11**, 1515–1544.

SIVAKUMARAN, S., AGAKOV, F., THEODORATOU, E., PRENDERGAST, J. G., ZGAGA, L., MANOLIO, T., RUDAN, I., MCKEIGUE, P., WILSON, J. F. AND CAMPBELL, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics* **89**, 607–618.

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. AND SMOLLER, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483–495.

TRAN, M.-N., NOTT, D. J., KUK, A. Y. C. AND KOHN, R. (2016). Parallel variational Bayes for large datasets with an application to generalized linear mixed models. *Journal of Computational and Graphical Statistics* **25**, 626–646.

WANG, N., GOSIK, K., LI, R., LINDSAY, B. AND WU, R. (2016). A block mixture model to map eQTLs for gene clustering and networking. *Scientific Reports* **6**, 21193.

XING, E. P., JORDAN, M. I. AND RUSSELL, S. (2002). A generalized mean-field algorithm for variational inference in exponential families. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann, pp. 583–591.

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

ZHOU, X. AND STEPHENS, M. (2014). Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature Methods* **11**, 407.

[*Received July 28, 2016; revised January 10, 2017; accepted for publication February 6, 2017*]