Bayesian Analysis (2012) **7**, Number 1, pp. 73–108

Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies

Peter Carbonetto* and Matthew Stephens†

Abstract. The Bayesian approach to variable selection in regression is a powerful tool for tackling many scientific problems. Inference for variable selection models is usually implemented using Markov chain Monte Carlo (MCMC). Because MCMC can impose a high computational cost in studies with a large number of variables, we assess an alternative to MCMC based on a simple variational approximation. Our aim is to retain useful features of Bayesian variable selection at a reduced cost. Using simulations designed to mimic genetic association studies, we show that this simple variational approximation yields posterior inferences in some settings that closely match exact values. In less restrictive (and more realistic) conditions, we show that posterior probabilities of inclusion for individual variables are often incorrect, but variational estimates of other useful quantities—including posterior distributions of the hyperparameters—are remarkably accurate. We illustrate how these results guide the use of variational inference for a genome-wide association study with thousands of samples and hundreds of thousands of variables.

Keywords: variable selection, variational inference, genetic association studies, Monte Carlo

1 Introduction

Many scientific questions are naturally framed as a variable selection problem: which variables $X_1, \ldots, X_p$ under investigation are useful for predicting outcome Y, assuming a linear model $E[Y] = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p$? Among the variety of approaches to variable selection for regression, the Bayesian approach (George and McCulloch 1997; Raftery, Madigan, and Hoeting 1997) stands out because we can assess the predictive value of a variable X_i simply by computing the posterior probability that it is included in the linear model (*i.e.* the posterior probability that its coefficient β_i is not zero). But exactly computing this posterior probability of inclusion is intractable because it involves summing over a combinatorially large number of models. Confronted with this fact, our goal is to make Bayesian variable selection viable for large problems with hundreds of thousands—if not millions—of variables that might explain outcome Y. We assess the potential of an approximation based on variational methods (Jordan et al. 1999) for achieving this aim.

The widespread use of the Bayesian approach to variable selection can be traced

*Department of Human Genetics, University of Chicago, Chicago, IL pcarbo@uchicago.edu
†Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL, mstephens@uchicago.edu

 DOI:10.1214/12-BA703

back to the advent of Markov chain Monte Carlo methods that effectively explore the posterior distribution (Clyde, Ghosh, and Littman 2011; Dellaportas, Forster, and Ntzoufras 2002; George and McCulloch 1993). MCMC methods avoid computing posterior probabilities for all 2^p combinations of predictors by focusing on subsets of high probability. But it can be difficult—or, at worst, prohibitive—to implement a Markov chain that efficiently explores the model space when we intend to investigate large numbers of variables that may predict Y. Our motivation is the analysis of genome-wide association studies (Servin and Stephens 2007; Stephens and Balding 2009); a present-day study can involve thousands of samples and hundreds of thousands of genetic variants that potentially explain a phenotype of interest (such as LDL cholesterol levels). Identifying the most promising genetic candidates could eventually point us to biological mechanisms underlying the phenotype. We would like to pursue the Bayesian approach to variable selection for genome-wide associations studies, and though sophisticated MCMC methods have been designed for this problem (Bottolo and Richardson 2010; Guan and Stephens 2011), they can take weeks to produce reasonably accurate inferences. And genetic association studies are only getting bigger—in the future we would like to tackle genome-wide association studies with millions of variables and hundreds of thousands of samples. The Lasso (Tibshirani 1996) and related penalized regression methods (Tibshirani 2011) that compute a posterior mode can more easily handle large variable selection problems, and in fact they have been applied to genome-wide association studies (He and Lin 2011; Hoggart et al. 2008; Wu et al. 2009). But these methods are less suited to the analysis of genetic association studies because, among other reasons, they do not easily quantify statistical support for individual associations (Guan and Stephens 2011).[1]

We investigate a two-part solution to this problem using variational methods (Jordan et al. 1999; Ormerod and Wand 2010; Wainwright and Jordan 2008) and importance sampling (e.g. Andrieu et al. 2003). Each part is straightforward to explain.

The basic idea behind the first part is to recast the problem of computing posterior probabilities—which is inherently a high-dimensional integration problem—as an optimization problem by introducing a class of approximating distributions, then optimizing some criterion to find the distribution within this class that best matches the posterior. To make this approach viable for large problems, we force the approximating distribution to observe a simple conditional independence property, following Logsdon, Hoffman, and Mezey (2010): each regression coefficient β_i is independent of the other regression coefficients *a posteriori*, given the observations and hyperparameters. (In the variational methods literature, this is known as a "mean field" approximation.) We then search for a distribution with this conditional independence property that fits the posterior as well as possible. This procedure scales linearly with the number of variables.

The second part to our solution is to use importance sampling to compute the low-dimension posterior of the hyperparameters. Since each importance weight includes the marginal likelihood of the hyperparameters, and since this marginal likelihood is in-

[1] Meinshausen et al. (2009) describe a way to derive p-values from Lasso estimates. But recent work by He and Lin (2011) suggests that this procedure may be too conservative for use in genome-wide association studies, and other high-dimensional variable selection problems.

tractable to compute, we replace it with a lower bound calculated using the variational approximation obtained in the first part. This idea of replacing the marginal likelihood with its variational lower bound is not new—for example, this is the idea behind variational expectation maximization, where the *maximum a posteriori* solution is replaced with the maximum of the lower bound (Blei et al. 2003; Heskes et al. 2004). This same idea is also used in several recent papers to improve variational inference (Bouchard and Zoeter 2009; Cseke and Heskes 2011; Ormerod 2011).

Variational estimates of posterior distributions can be inaccurate. For instance, they are often too concentrated. In some cases this inaccuracy is not a problem, such as when the goal is prediction or point estimation. But in genome-wide association studies accurate computation of posterior probabilities is important because reports of new genetic associations for disease may lead to substantial investment in follow-up studies, and so they are received with a high level of scrutiny. For this reason, we focus on assessing the accuracy of the variational approximation. To be clear, we are concerned with accuracy of the approximate computations, not accuracy of the predictions. In our motivating problem, most genetic loci will be unlinked because they are on separate chromosomes, or they will be weakly linked because of recombination. Therefore, X_i and X_j will be nearly independent for most pairs i and j. (For our choice of prior, independence of X_i and X_j implies near independence of their effects on Y under the posterior, as we explain below.) In this case, the variational approximation recovers accurate posterior inclusion probabilities and other quantities of interest. In situations where the conditional independence assumption is violated, we would not expect accurate approximations of the posterior inclusion probabilities. And yet, we show that the variational method can provide useful inferences in these cases—including accurate posterior distributions of the hyperparameters—even when the posterior inclusion probabilities are incorrect.

Our method builds on a variational approximation recently developed in the same context (Logsdon et al. 2010). (It is also closely related to the approximating distribution developed for independent factor analysis in Attias 1999.) The principal difference between our method and theirs is that their method imposes independence assumptions on the hyperparameters, whereas ours does not. Instead, we use importance sampling to compute the posterior distribution of the hyperparameters. Since these additional independence assumptions do not hold in general, avoiding them seems preferable. In addition, there are some differences in our model and priors; for example, Logsdon et al. (2010) have separate prior distributions for positive and negative effects, whereas we do not. Most importantly, the emphasis of our paper is very different: we focus on the accuracy of the variational approximation compared to exact or MCMC-based calculations.

To validate the variational approximation, we present two simulation studies (Sec. 5): an idealized simulation in which all variables are independent, and a more realistic case study in which many variables are strongly correlated. These simulation studies are small enough that we can assess the accuracy of our answers by comparing them to Monte Carlo computations. We also illustrate the features (and possible issues) of the variational approximation with a small example in Sec. 4.

Informed by the results of our simulation studies, in Sec. 6 we demonstrate the use of our variational inference procedure in a case-control study to identify genetic factors that contribute to a complex human disease. We complete the full analysis of ~ 400,000 genetic variants and ~ 5,000 samples in hours, an analysis that might otherwise take days or weeks by simulating a Markov chain.

In Sec. 2, we describe the hierarchical model for variable selection, assuming a linear model for Y. In Sec. 3, we present the technical details of our inference procedure for Bayesian variable selection. For binary outcomes in the case-control study, we describe an extension to our inference method, the details of which are given in the appendix.

2 Bayesian variable selection: background and notation

There are many possible approaches to Bayesian variable selection; see O'Hara and Sillanpää (2009) for a recent review. The focus of this paper is an approach based on a sparse ("spike and slab") prior for the coefficients of the linear regression. This is one of the most widely used approaches to Bayesian variable selection in linear regression. The sparsity of the prior has a particular appeal for genetic association studies where most genetic variants have no effect on the outcome, or an effect that is indistinguishable from zero, even in large samples. The variational inference procedure we describe does not exploit the sparsity of the prior, so it may be possible to extend it to non-sparse priors that induce shrinkage in the regression coefficients, including normal-gamma priors (Griffin and Brown 2010) and the Bayesian Lasso (Park and Casella 2008), but investigating this question lies outside the scope of this paper.

Following standard practice, we model the variable of interest Y as a linear combination of the candidate predictors $X = (X_1, \ldots, X_p)^T$ plus residual noise $\varepsilon \sim N(0, \sigma^2)$:

$$Y = \beta_0 + \sum_{k=1}^{p} X_k \beta_k + \varepsilon. \tag{1}$$

The variable selection problem can be viewed as deciding which of the coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$ are equal to zero. We use binary variables $\gamma = (\gamma_1, \ldots, \gamma_p)^T$ to indicate whether or not each variable is included in the model; if $\gamma_k = 0$, then $\beta_k = 0$ with probability 1. Pursuing a Bayesian approach to variable selection, we assign priors to the indicator variables γ and coefficients β, then compute posterior probabilities by averaging over choices of β and γ, and any additional model parameters, such as σ^2.

In our problem formulation, the data consist of an $n \times p$ matrix $\mathbf{X}$ of observations on the independent variables X, and a vector $y = (y_1, \ldots, y_n)^T$ of observed values of Y. We account for an intercept β_0 in the linear model (1) by centering y and the columns of $\mathbf{X}$ so that they each have a mean of zero. This is equivalent to integrating out the intercept with respect to an improper, uniform prior (Chipman, George, and McCulloch 2001).[2] The extension to binary labels $Y \in \{0, 1\}$, which is needed for the case-control study of Sec. 6, is covered in the appendix.

[2] In general, one must be careful with the use of improper priors in variable selection (Clyde and

There are many ways to specify a prior on subsets. For simplicity, we assume an exchangeable prior, and treat indicator variables γ as Bernoulli random trials governed by common success rate $p(\gamma_k = 1) = \pi$. This is the "spike and slab" prior (Mitchell and Beauchamp 1988), in which β_k is drawn from the "slab" density with probability π and, with probability $1 - \pi$, β_k equals zero (the "spike"). We take the slab density to be normal with zero mean and variance $\sigma^2\sigma_\beta^2$. In many applications π will be small, reflecting a low proportion of variables with nonzero coefficients.

Statisticians may have good reasons to prefer other priors for β and γ, and our variational approximation could easily accommodate other priors, including nonexchangeable priors for γ, and the conventional g-prior for β (Liang et al. 2008; Zellner 1986). We discuss this point below.

Since results can be sensitive to the choice of hyperparameters $\theta = (\sigma^2, \sigma_\beta^2, \pi)$, we estimate θ from the data by introducing a prior on θ, and integrating over values of θ. We do not assume a specific form for the prior on θ—one feature of our variational method is that it works with any prior on the hyperparameters. We defer the choice of prior to the experiments (see Sections 5 and 6).

The inference problem is to compute posterior probabilities, or expected values with respect to the posterior. For example, the posterior probability that variable X_k is included in the linear model of Y is

$$\mathrm{PIP}(k) \equiv p(\gamma_k = 1 \,|\, \mathbf{X}, y) = \frac{\sum_{\gamma_{-k}} \int\int p(y, \beta, \gamma_k = 1, \gamma_{-k} \,|\, \mathbf{X}, \theta)\, p(\theta)\, d\beta\, d\theta}{\sum_{\gamma} \int\int p(y, \beta, \gamma \,|\, \mathbf{X}, \theta)\, p(\theta)\, d\beta\, d\theta}, \tag{2}$$

where γ_{-k} is an assignment to all the indicator variables except γ_k. Since we refer to this probability often, we abbreviate it as PIP, for "posterior inclusion probability." The joint probability of y, β and γ given $\mathbf{X}$ and θ is

$$p(y, \beta, \gamma \,|\, \mathbf{X}, \theta) = p(y \,|\, \mathbf{X}, \beta, \sigma^2) \prod_{k=1}^{p} p(\beta_k \,|\, \gamma_k, \sigma^2, \sigma_\beta^2) \prod_{k=1}^{p} p(\gamma_k \,|\, \pi). \tag{3}$$

The posterior inclusion probability contains a sum over 2^p possible models γ, an integral of high dimension over the nonzero coefficients β, and an additional integral over the hyperparameters θ. MCMC methods approximate the intractable sums and integrals by implementing Metropolis-Hastings moves that explore models with strong support under the posterior. The challenge lies in designing a Markov chain that explores the model space efficiently, and a variety of ways to deal with this issue have been suggested; see Bottolo and Richardson (2010), Clyde et al. (2011), and Dellaportas et al. (2002) for overviews. We investigate an alternative approach using variational methods.

George 2004). However, in this case the improper priors we use for β_0, and for σ^2 later, result in well-defined Bayes factors and posterior probabilities (Servin and Stephens 2007).

- **Inputs:** $\mathbf{X}$, y, $\theta^{(1)}, \ldots, \theta^{(N)}$, $\tilde{p}(\theta^{(1)}), \ldots, \tilde{p}(\theta^{(N)})$.
- **Outputs:** $\hat{\alpha}$, $\hat{\mu}$, $w(\theta^{(1)}), \ldots, w(\theta^{(N)})$.
- Choose $\alpha^{(\mathsf{init})}$ and $\mu^{(\mathsf{init})}$.
- **for** $i = 1, \ldots, N$ ***(outer loop)***
  - Set $\alpha = \alpha^{(\mathsf{init})}$ and $\mu = \mu^{(\mathsf{init})}$.
  - Set $\theta = \theta^{(i)}$.
  - Solve for s^2; see (8).
  - **Repeat** until convergence ***(inner loop)***
    1. Choose $k \in \{1, \ldots, p\}$.
    2. Update μ_k according to (9).
    3. Update α_k according to (10).
  - Set Z to the lower bound on $p(y \,|\, \mathbf{X}, \theta)$; see (14).
  - Compute unnormalized importance weight: set $w(\theta^{(i)}) = Z/\tilde{p}(\theta^{(i)})$.
  - Set $\alpha^{(i)} = \alpha$ and $\mu^{(i)} = \mu$.
- Compute normalized importance weights $\hat{w}(\theta^{(i)})$.
- Average over hyperparameters:
  - set $\hat{\alpha} = \hat{w}(\theta^{(1)})\,\alpha^{(1)} + \cdots + \hat{w}(\theta^{(N)})\,\alpha^{(N)}$.
  - set $\hat{\mu} = \hat{w}(\theta^{(1)})\,\mu^{(1)} + \cdots + \hat{w}(\theta^{(N)})\,\mu^{(N)}$.

Figure 1: Outline of inference procedure for Bayesian variable selection. The input arguments are the samples $(\mathbf{X}, y)$, and the hyperparameter values $\theta^{(1)}, \ldots, \theta^{(N)}$ drawn from importance sampling distribution $\tilde{p}(\theta)$. The outputs are normalized importance weights $\hat{w}(\theta^{(i)}) \approx p(\theta^{(i)} \,|\, \mathbf{X}, y)$, and posterior probabilities $\hat{\alpha}_k \approx \mathrm{PIP}(k)$ and mean additive effects $\hat{\mu} \approx E[\beta_k \,|\, \gamma_k = 1]$ averaged over settings of the hyperparameters. In practice, we run a separate optimization to choose $\alpha^{(\mathsf{init})}$ and $\mu^{(\mathsf{init})}$. This is done to address convergence of the inner loop to local maxima (see Sec. 3.2).

3 Variational inference

We begin by decomposing the posterior inclusion probabilities as

$$\mathrm{PIP}(k) = \int p(\gamma_k = 1 \,|\, \mathbf{X}, y, \theta)\, p(\theta \,|\, \mathbf{X}, y)\, d\theta. \tag{4}$$

There are two components to our our inference strategy. One component approximates posterior probabilities $p(\gamma_k = 1 \,|\, \mathbf{X}, y, \theta)$ by minimizing the Kullback-Leibler divergence (Cover and Thomas 2006) between an approximating distribution on β, γ and the posterior of β, γ given θ. The second component estimates $p(\theta \,|\, \mathbf{X}, y)$ by importance sampling, using the variational solution from the first component to compute the importance weights. The final inference procedure is shown in Fig. 1: the first component is the inner loop, and the second component is the outer loop of the algorithm.

3.1 Posterior inclusion probabilities given hyperparameters

The inner loop searches for a distribution $q(\beta, \gamma)$ that provides a good approximation to the posterior $f(\beta, \gamma) = p(\beta, \gamma \,|\, \mathbf{X}, y, \theta)$. This is accomplished by minimizing the Kullback-Leibler divergence

$$D(q \,\|\, f) = \int q(\beta, \gamma) \log \big\{ q(\beta, \gamma) / f(\beta, \gamma) \big\} \, d\beta \, d\gamma. \tag{5}$$

We restrict $q(\beta, \gamma)$ to be of the form

$$q(\beta, \gamma; \phi) = \prod_{k=1}^{p} q(\beta_k, \gamma_k; \phi_k). \tag{6}$$

where $\phi = (\phi_1, \ldots, \phi_p)$ are free parameters, and the individual factors have the form

$$q(\beta_k, \gamma_k; \phi_k) = \begin{cases} \alpha_k N(\beta_k \,|\, \mu_k, s_k^2) & \text{if } \gamma_k = 1; \\ (1 - \alpha_k) \, \delta_0(\beta_k) & \text{otherwise,} \end{cases} \tag{7}$$

where $\delta_0(\,\cdot\,)$ is the delta mass (or "spike") at zero, and $\phi_k = (\alpha_k, \mu_k, s_k^2)$. With probability α_k, the additive effect β_k is normal with mean μ_k and variance s_k^2 (the "slab"), and with probability $1 - \alpha_k$, the variable has no effect on Y.

This "fully-factorized" approximating distribution was first suggested by Logsdon et al. (2010), and Attias (1999) proposed it for a related model. It can be motivated by the observation that, under the priors we adopt here, the posterior of β and γ will be of this form when $\mathbf{X}^T\mathbf{X}$ is diagonal. Of course, it is unreasonable to expect that each off-diagonal entry $(\mathbf{X}^T\mathbf{X})_{jk}$ is exactly zero. But if variables X_j and X_k are independent, and if the expected value of X_j and X_k is zero—which is guaranteed once we center the columns of $\mathbf{X}$—then $(\mathbf{X}^T\mathbf{X})_{jk}$ will be close to zero, and β_j and β_k will be nearly independent *a posteriori* given the additive effects of the remaining variables. Therefore, we expect that (6) will be a good approximation when the variables are independent. It will also be a good approximation when the posterior is concentrated at a single location. Note that these arguments would be equally valid if we instead used the g-prior for β.

Finding the best fully-factorized distribution $q(\beta, \gamma; \phi)$ amounts to finding the free parameters ϕ that make the Kullback-Leibler divergence as small as possible. The coordinate descent updates for this optimization problem can be obtained by taking partial derivatives of the Kullback-Leibler divergence, setting the partial derivatives to zero, and solving for the parameters α_k, μ_k and s_k^2. This yields coordinate updates

$$\mathbb{V}\mathrm{ar}[\beta_k \,|\, \gamma_k = 1] \;\approx\; s_k^2 \;=\; \frac{\sigma^2}{(\mathbf{X}^T\mathbf{X})_{kk} + 1/\sigma_\beta^2} \tag{8}$$

$$E[\beta_k \,|\, \gamma_k = 1] \;\approx\; \mu_k \;=\; \frac{s_k^2}{\sigma^2} \Big((\mathbf{X}^T y)_k - \sum_{j \neq k} (\mathbf{X}^T\mathbf{X})_{jk} \alpha_j \mu_j \Big) \tag{9}$$

$$\frac{p(\gamma_k = 1 \,|\, \mathbf{X}, y, \theta)}{p(\gamma_k = 0 \,|\, \mathbf{X}, y, \theta)} \;\approx\; \frac{\alpha_k}{1 - \alpha_k} \;=\; \frac{\pi}{1 - \pi} \times \frac{s_k}{\sigma_\beta \sigma} \times e^{\mathrm{SSR}_k/2}, \tag{10}$$

where $(\mathbf{X}^T y)_k$ is the kth entry of vector $\mathbf{X}^T y$, and $\mathrm{SSR}_k = \mu_k^2/s_k^2$. Note that α_k, μ_k and s_k^2 all implicitly depend on the value of θ. The inner loop of the inference algorithm repeatedly applies updates (8-10) until a stationary point is reached.

Expressions (8) and (10) may look familiar: (8) is the posterior variance of the additive effect β_k for the single-variable linear model $Y = X_k\beta_k + \varepsilon$; and (10) is the posterior odds (Bayes factor $\times$ prior odds) for the alternative hypothesis ($\beta_k \neq 0$) over the null hypothesis ($\beta_k = 0$), assuming that μ_k is the correct posterior mean, in which case SSR_k is the reduction in sum of squares due to regression on X_k.

Likewise, (9) is also easy to explain: if we ignore all terms involving variables j other than variable k, it is the posterior expected value of β_k for the single-variable linear model $Y = X_k\beta_k + \varepsilon$. The $(\mathbf{X}^T\mathbf{X})_{jk}\alpha_j\mu_j$ terms correct for correlations among variables not included in the single-variable linear model. For example, when another variable X_j is positively correlated with X_k, and we already know it has an effect on Y in the same direction as X_k, equation (9) dampens the effect of X_k on Y. This correction also accounts for the probability that variable X_j is included in the model.

One final comment on the first part of our inference procedure: the algorithm as we present it in Fig. 1 does not scale linearly with the number of variables. The most expensive part is the update for μ_k. The trick to implementing inner loop iterations with linear complexity is to keep track of vector $\mathbf{X}r$, where r is a column vector with entries $r_k = \alpha_k\mu_k$, and to update this vector after each update of μ_k and α_k.

3.2 Posterior of hyperparameters

We use importance sampling to integrate over the hyperparameters. We replace integral (4) with importance sampling estimate

$$\mathrm{PIP}(k) \approx \frac{\sum_{i=1}^{N} p(\gamma_k = 1 \mid \mathbf{X}, y, \theta^{(i)})\, w(\theta^{(i)})}{\sum_{i=1}^{N} w(\theta^{(i)})}, \tag{11}$$

where $w(\theta)$ is the unnormalized importance weight for θ. Other Monte Carlo methods such as MCMC could also be used to integrate over the hyperparameters, but we opt for importance sampling because it is a simple and effective way to estimate an integral of low dimension, and because we can obtain a reasonably accurate estimate with a small number of samples, provided they are chosen well. This is an important consideration because a single iteration of importance sampling involves optimizing a variational lower bound (as we explain below) and this can take a long time to complete for large problems. In our analyses, we use a small number of samples of θ, between 100 and 1000.

By replacing integral (4) with the Monte Carlo estimate (11), we avoid having to introduce additional variational approximations for the hyperparameters. The difficulty, however, is that the importance weights are

$$w(\theta) = \frac{p(y \mid \mathbf{X}, \theta)\, p(\theta)}{\tilde{p}(\theta)}, \tag{12}$$

where $\tilde{p}(\theta)$ is the importance sampling distribution; this expression contains a marginal likelihood $p(y \mid \mathbf{X}, \theta)$ which we know by now is difficult to compute. We take a variational-based approach to approximating this importance weight.

Our approach is based on the previously established result that the marginal log-likelihood of θ is bounded from below by

$$\log p(y \mid \mathbf{X}, \theta) \geq F(\theta; \phi) \equiv \int\int q(\beta, \gamma; \phi) \log \left\{ \frac{p(y, \beta, \gamma \mid \mathbf{X}, \theta)}{q(\beta, \gamma; \phi)} \right\} d\beta \, d\gamma. \tag{13}$$

For our choice of approximating distribution, this lower bound has analytical expression

$$\begin{aligned} F(\theta; \phi) = &-\frac{n}{2} \log(2\pi\sigma^2) - \frac{\|y - \mathbf{X}r\|^2}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{k=1}^{p} (\mathbf{X}^T\mathbf{X})_{kk} \mathbb{V}\mathrm{ar}[\beta_k] \\ &- \sum_{k=1}^{p} \alpha_k \log\left(\frac{\alpha_k}{\pi}\right) - \sum_{k=1}^{p} (1 - \alpha_k) \log\left(\frac{1 - \alpha_k}{1 - \pi}\right) \\ &+ \sum_{k=1}^{p} \frac{\alpha_k}{2} \left[1 + \log\left(\frac{s_k^2}{\sigma_\beta^2 \sigma^2}\right) - \frac{s_k^2 + \mu_k^2}{\sigma_\beta^2 \sigma^2}\right], \end{aligned} \tag{14}$$

where $\|\cdot\|$ is the Euclidean norm, and $\mathbb{V}\mathrm{ar}[\beta_k] = \alpha_k(s_k^2 + \mu_k^2) - (\alpha_k\mu_k)^2$ is the variance of kth additive effect under the approximating distribution. This bound is valid for any θ and ϕ. See Jordan et al. (1999) for a derivation of this bound using Jensen's inequality.

It is easy to see that the minimizer of the Kullback-Leibler divergence for a given hyperparameter setting θ, which we denote by $\phi(\theta)$, also maximizes the lower bound $F(\theta; \phi)$. In other words, $\phi(\theta)$ provides the tightest lower bound—hence the best approximation to the marginal likelihood—within a particular family of approximating distributions. Motivated by this, others (e.g. Blei et al. 2003; Khan et al. 2010) have proposed to replace the intractable maximum likelihood estimator for θ with a θ that maximizes the best lower bound, $F(\theta; \phi(\theta))$. Likewise, we propose to substitute the marginal log-likelihood appearing in the importance weight (12) with its corresponding best lower bound, $F(\theta, \phi(\theta))$.[3]

In general, there is no reason to believe that $F(\theta; \phi(\theta))$ is a good substitute for the marginal log-likelihood. In fact, it is often a poor substitute, as we show in the examples below. However, all that is needed for our inference procedure to work well is that $F(\theta; \phi(\theta))$ have a similar shape to $\log p(y \mid \mathbf{X}, \theta)$ whenever the marginal likelihood is relatively large. By the same logic, computing θ that maximizes $F(\theta; \phi(\theta))$ is sensible so long as the maximum of the lower bound is close to the maximum likelihood estimate.

[3]In our experiments, we choose samples $\theta^{(i)}$ on a fixed grid to reduce the variance in the Monte Carlo estimates. This is feasible since we only have 2 or 3 hyperparameters. In this case, our inference strategy resembles "grid-based" variational inference (Cseke and Heskes 2011; Ormerod 2011), but there is an important difference: we treat θ differently from the other variables (β, γ) because we never use variational inference to compute an integral over θ. Our method is more accurate, but more costly, because we need to re-run the variational inference portion (the "inner loop") separately for each $\theta^{(i)}$.

The problem is that there are no theoretical results guaranteeing the accuracy of estimates based on the lower bound (13), and in most applications it seems to be simply taken on faith. In this paper, we assess the accuracy of the approximation empirically by comparing variational estimates to exact calculations (or MCMC calculations where exact calculations are infeasible). In our experiments, the resulting approximate importance weights are often very accurate.

We run the coordinate ascent updates separately for each setting of the hyperparameters, with common starting point $(\alpha^{(\mathsf{init})}, \mu^{(\mathsf{init})})$. Since the coordinate ascent updates are only guaranteed to converge to a local minimum of the Kullback-Leibler divergence, the choice of starting point can affect the quality of the approximation, particularly when variables are correlated. To address sensitivity of the approximation to local maxima, we select a common starting point by first running the inner loop for each $\theta^{(i)}$, with random initializations for α and μ, then we assign $(\alpha^{(\mathsf{init})}, \mu^{(\mathsf{init})})$ to the solution $(\alpha^{(i)}, \mu^{(i)})$ from the hyperparameter setting $\theta^{(i)}$ with the largest marginal likelihood.

4 An illustration

In this section, we compare variational estimates of posterior distributions with exact calculations in a small variable selection problem with two candidate predictors.

The problem setup is as follows. We take Y to be a linear combination of the variables, $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$, with random error ε drawn from the standard normal. If variable X_k is included in the model, then $\beta_k \neq 0$. Since there are only four possible models or combinations of included variables to choose from, it is easy to compute posterior probabilities of all models $\gamma \in \{0,1\} \times \{0,1\}$. Each posterior probability $p(\gamma \,|\, \mathbf{X}, y)$ is computed by averaging over nonzero coefficients β. We place a normal prior on $\beta_1 \,|\, \gamma_1 = 1$ and $\beta_2 \,|\, \gamma_2 = 1$ with mean zero and standard deviation 0.1, and an improper, uniform prior on intercept β_0. Each γ_k is *i.i.d.* Bernoulli with success rate π. The data are $n = 1000$ samples of X_1, X_2 and Y.

To make this example more interesting, we treat π as unknown. The posterior probability of any γ is then averaged over choices of π. We take π to be Beta$(0.2, 2)$. Note that this prior favours sparse models; it says that, in expectation, only 1 out of 10 variables are included in the model.

Our first example, Example A in Fig. 2, is designed to illustrate a setting where the variational approximation should perform well in all aspects, because the two variables X_1 and X_2 are only weakly correlated, with correlation coefficient $r = 0.2$. The first variable has a modest effect on Y; the coefficients used to simulate the samples y are $(\beta_0, \beta_1, \beta_2) = (0, 0.1, 0)$. Observe that the posterior inclusion probabilities shown in Fig. 2 correctly favour X_1 as a predictor of Y. Since the variables are weakly correlated, we have reason to expect that the fully-factorized distribution will be a good fit to the posterior. Indeed, this is what we observe: the posterior probabilities and marginal likelihoods under the variational approximation (in gray) all closely match exact calculations (in black).

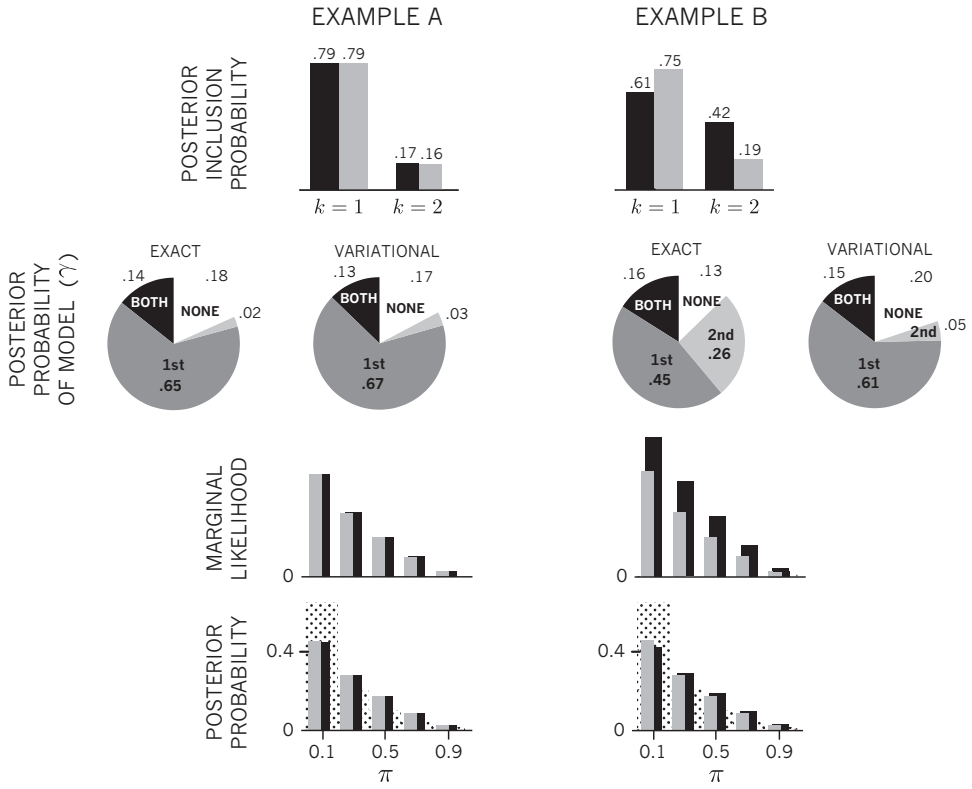

Figure 2: Two toy examples illustrating some of the features of the variational approximation. In bar plots, variational estimates are gray and exact computations are black. Note that the scale of the marginal likelihoods does not matter, only their relative values do. In the bottom row, the prior on π is drawn in a stippled pattern. See the text for details about each example.

The second example (Example B) is designed to illustrate a less ideal situation for the variational approach where the variables are more strongly correlated; $r = 0.8$. (The true coefficients remain the same.) The posterior shown in Fig. 2 still favours X_1 over X_2, but with less certainty because of the higher correlation. Due to the correlation between X_1 and X_2, we no longer expect that the fully-factorized distribution will correctly capture the posterior. This suspicion is correct: the variational approximation overestimates the posterior probability that X_1 is included in the model, and underestimates the posterior inclusion probability for X_2. The tendency to concentrate more mass on a single hypothesis, or to artificially lower the variance in the posterior by overly favouring the winner, is typical behaviour of mean field approximations (MacKay 2003; Turner et al. 2008). The fully-factorized approximation cannot capture the posterior distribution over models because, for example, $p(\gamma_1 = 1, \gamma_2 = 1 \,|\, \mathbf{X}, y) = 0.16$ cannot be written as the product of $p(\gamma_1 = 1 \,|\, \mathbf{X}, y) = 0.61$ and $p(\gamma_2 = 1 \,|\, \mathbf{X}, y) = 0.42$.

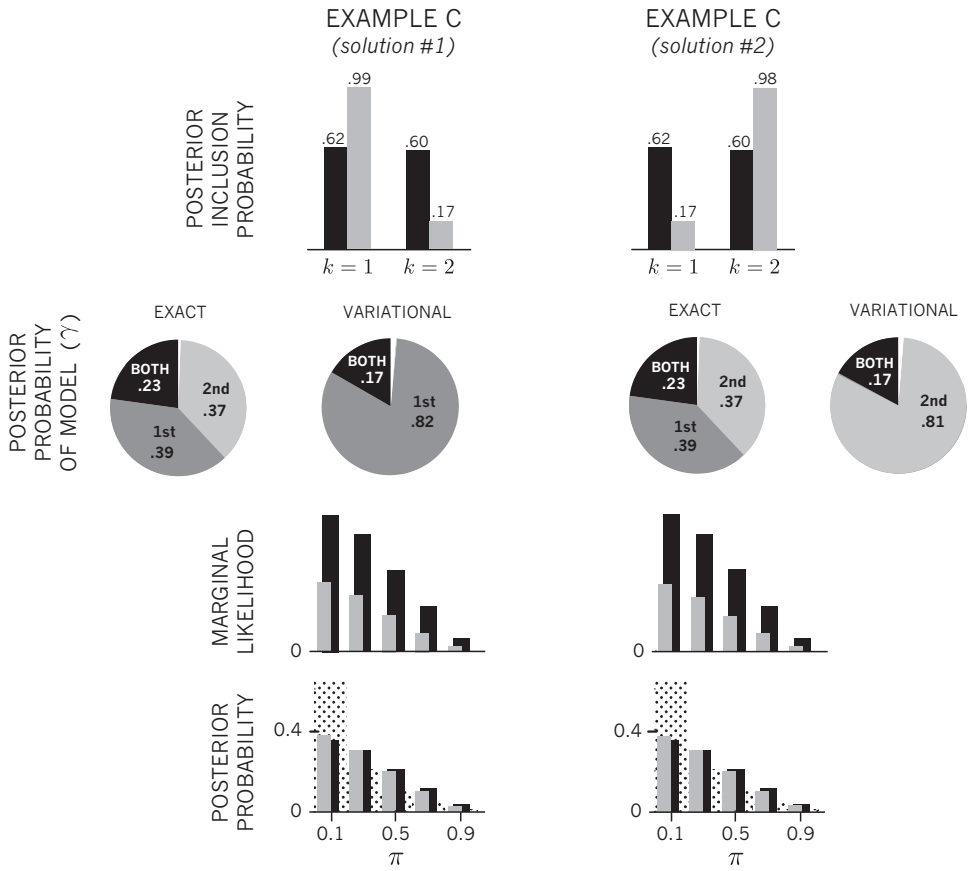

Figure 3: Another toy example demonstrating some features of the variational approximation. In bar plots, variational estimates are shown in gray, and exact computations are black. The left and right columns show the two variational solutions. (Exact computations remain the same in both columns.) In the bottom row, the prior on π is drawn in a stippled pattern.

Despite this limitation, the variational approximation still provides a good estimate for the posterior of π (bottom row of Fig. 2). This is observed even though the variational lower bound $F(\pi; \phi(\pi))$ (third row) is a poor approximation to the marginal likelihood $p(y \,|\, \mathbf{X}, \pi)$. But since $F(\pi; \phi(\pi))$ has a similar shape to the marginal likelihood, we obtain the correct posterior $p(\pi \,|\, \mathbf{X}, y)$ after normalizing. Consistent with this result, the variational approximation also provides an accurate posterior distribution for the number of variables included in the model.

The third example, Example C in Fig. 3, is intended to illustrate the behaviour of the variational approximation when the two variables are almost completely correlated ($r = 0.99$), in which case it is difficult to distinguish the first variable (which affects

Y) from the second (which does not). Indeed, the posterior inclusion probabilities are 0.62 and 0.60. In this case, there are two local maxima for the free parameters ϕ which produce very different approximations to the posterior inclusion probabilities, and both these approximations yield poor estimates of the posterior inclusion probabilities. Nonetheless, as in Example B, both solutions provide accurate posterior distributions for π and the number of variables included (the largest error in both instances is 0.06), despite the fact that the variational lower bound drastically underestimates the marginal likelihood.

Of course, for most larger problems our variational approximation will be inadequate for capturing complicated dependencies among the variables, and the estimates of the posterior will suffer accordingly. When a more precise answer is needed, MCMC may be the better, if more costly, option because it (eventually) averages over all credible models. The goal of these examples was to point out that the variational method can often do a good job estimating some posterior quantities (such as π and the number of included variables), even if it fails to capture the multi-modality of the posterior, by choosing models that are reasonably representative of the full range of possibilities. If accurate probabilities for individual variables are not critical, the variational method can be an adequate and much less costly option.

5 Two simulation studies

Now we present two simulation studies to assess the accuracy of the variational approximation for variable selection. The first experiment is an idealized genetic association study with uncorrelated genetic factors. The second experiment represents a situation in which we target a specific region of the genome, and we have sampled genetic variants in that region. In the second case, many genetic factors are strongly correlated.

5.1 The ideal case

Earlier, we argued that the variational method should yield accurate posterior inclusion probabilities when the variables are independent. The purpose of our first experiment is to assess this claim. The variables for this experiment are modeled after genetic variants—specifically, single-nucleotide polymorphisms (SNPs).

In a typical genome-wide association study, most genetic variants do not contribute to changes in the quantitative trait Y, so the inferred β should be sparse. Moreover, the accumulated effect of genetic factors usually only accounts for a modest portion of variance in the trait. This can be due to a variety of reasons: we failed to measure some of the variants that affect Y, such as structural variants; there are other factors, such as environmental factors, that play a role in determining Y; and perhaps there are interactions among genetic factors that cannot be captured by a linear model.

To generate the genotype data $\mathbf{X}$ for our experiment, we start by selecting, for each SNP $k = 1, \ldots, p$, the frequency f_k that its minor allele appears in the population. We

	$E[\log_{10}\sigma^2]$	$E[\log_{10}\sigma_\beta^2]$	$E[\log_{10}\pi]$
variational	0.954	−0.803	−1.86
MCMC	0.939	−0.860	−1.76
difference	0.015	0.057	−0.10

Table 1: Posterior means of $\log_{10}$ hyperparameters for a typical trial from the first simulation study with independent SNPs. The top two rows show variational and MCMC estimates of posterior expected values. The bottom row shows differences $\tilde{x} - x$, where x is the MCMC estimate and $\tilde{x}$ is the variational approximation.

sample minor allele frequencies f_k *i.i.d.* from the uniform distribution on $[0.05, 0.5]$. This is intended to mimic a genome-wide association study with "common" genetic variants. (The distribution of minor allele frequencies is not uniform in some more recent studies because genotyping platforms now have better coverage of rare variants.) Then for each SNP k and individual i we simulate the genotype x_{ik} independently from the binomial for two trials (corresponding to the two alleles) and with success rate f_k.

We then generate genotypes $y = (y_1, \ldots, y_n)^T$ for n individuals. To do so, we first select m SNPs uniformly at random to have non-zero coefficients β_k, and sample these coefficients *i.i.d.* from the standard normal. Then we set $y_i = \sum_{k=1}^{p} x_{ik}\beta_k + \varepsilon_i$, where the error terms ε_i are *i.i.d* from $N(0, \sigma^2)$.

We repeat this process of generating SNPs and samples 50 times to generate data sets for 50 separate experimental trials. For all trials, we set $n = 500$, $p = 1000$, $m = 20$ and $\sigma = 3$. While these settings lead to data sets that are much smaller than real genome-wide association studies, they capture some of their key characteristics—the true model is sparse, genetic factors explain on average about half the variance in Y—while producing data sets small enough that we can run many simulations in a reasonable amount of time.

We implement Bayesian variable selection as it was described in Sec. 2. We follow a hierarchical Bayesian strategy, specifying priors for the hyperparameters $\theta = (\sigma^2, \sigma_\beta^2, \pi)$, and estimating their posterior distribution from the data. We adopt the standard prior $p(\sigma^2) \propto 1/\sigma^2$ for the residual variance parameter (Berger 1985), and a Beta$(0.02, 1)$ prior for π. The prior for π has mean equal to $20/1000$, which is the true proportion of variables that affect Y. However, the prior is diffuse, and is skewed toward small models; for example, the prior probability that more than one variable is included in the model is 0.07. This prior may not be appropriate for general application to genetic association studies, but we use it here to facilitate implementation of the MCMC method, as the beta prior allows us to analytically integrate out π. In our case study (Sec. 6), we switch to a normal prior on $\log \frac{\pi}{1-\pi}$.

For the prior variance parameter σ_β^2, we adopt a prior related to the one recommended by Guan and Stephens (2011). Based on arguments given in Guan and Stephens (2011), it is appropriate to place a prior on the expected proportion of variance explained $\hat{r}^2 = \hat{s}_z^2/(1 + \hat{s}_z^2)$, where $\hat{s}_z^2 = \pi\sigma_\beta^2 \sum_{k=1}^{p} \hat{s}_k^2$, and where $\hat{s}_k^2$ is the sample variance of the

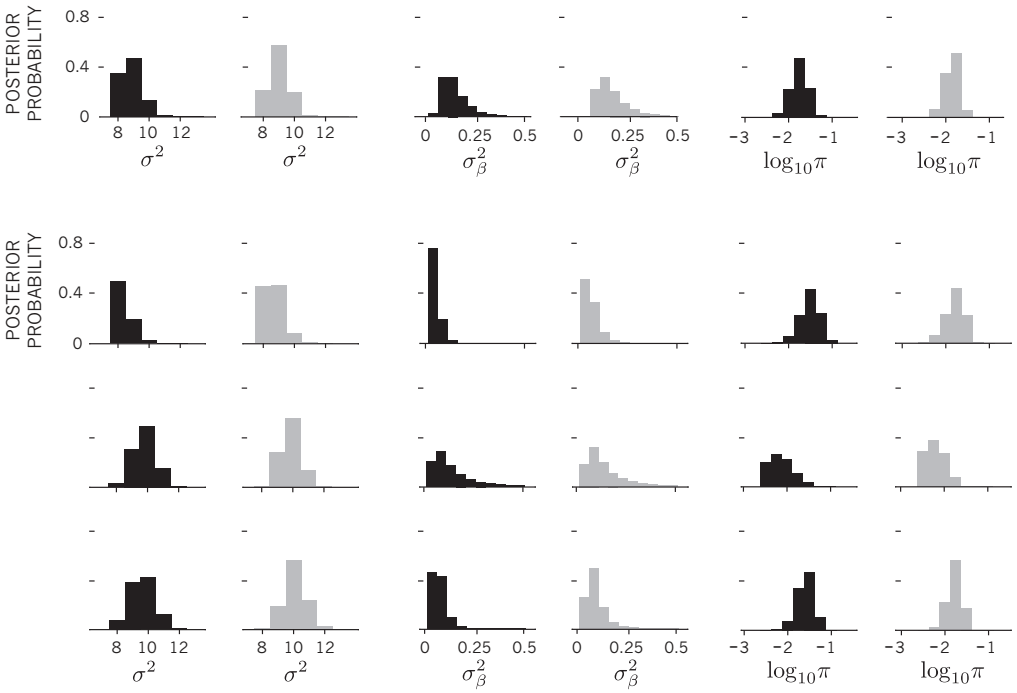

Figure 4: Posterior of hyperparameters for several trials in idealized simulation study with independent SNPs. Variational estimates are gray, and MCMC computations are black. Each row corresponds a single trial. The top row is a trial demonstrating typical behaviour. The other rows are outliers; specifically, trials exhibiting largest differences between variational and MCMC estimates of posterior means of the hyperparameters.

kth variable. (Note that $\hat{s}_z^2$ times the scale parameter σ^2 is the prior expected value of the sample variance of $X^T\beta$.) This leads to a diffuse (heavy-tailed) prior on σ_β^2 that depends on π. While the prior dependence of σ_β^2 and π is useful, for convenience of implementing MCMC we avoid this dependence by replacing π with a constant, 0.02, that represents its true value.

Our variational inference method requires specification of an importance sampling distribution on θ. We take $\tilde{p}(\theta)$ to be uniform on $(\sigma^2, \sigma_\beta^2, \log_{10}\pi)$, and we set the range of the uniform distribution to be sufficiently large to include all values with appreciable posterior probability. (Defining $\tilde{p}(\theta)$ on a wider range would not change the final results, and would increase the running time of the experiments.) To reduce the variance of importance sampling, rather than actually sampling from the uniform proposal distribution, we use a deterministic, regular grid of values for $\theta^{(i)}$. Values of σ^2, σ_β^2 and $\log_{10}\pi$ are taken at regular intervals of 1, 0.025 and 0.25, respectively. These intervals were chosen after some trial and error to produce approximately the same resolution of the posterior distribution in each dimension.

	$E[\log_{10}\sigma^2]$	$E[\log_{10}\sigma_\beta^2]$	$E[\log_{10}\pi]$
variational	0.98 ± 0.06	-0.89 ± 0.31	-1.85 ± 0.27
MCMC	0.96 ± 0.07	-0.93 ± 0.36	-1.77 ± 0.33
mean diff.	0.013	0.042	-0.087
mean abs. diff.	0.013	0.048	0.089

Table 2: Posterior mean estimates of $\log_{10}$ hyperparameters from the first simulation study, averaged over all 50 trials. Standard error ($\pm$) is two times the sample deviation over the 50 trials. The top two rows show variational and MCMC estimates of the posterior expected values. The third and fourth rows show the mean of differences $\tilde{x} - x$, and the mean of absolute differences $|\tilde{x} - x|$, where x is the MCMC estimate and $\tilde{x}$ is the variational approximation.

To assess the accuracy of the inferences provided by the variational approximation, we compare the results from the variational method to Monte Carlo estimates of posterior distributions obtained from running an MCMC algorithm for 100,000 iterations (see the appendix for details). We cannot, of course, guarantee that 100,000 iterations of MCMC, or any finite number of iterations, is sufficient to recover accurate posterior quantities, but since we cannot calculate exact posterior probabilities we must tolerate some degree of imprecision in our evaluation.

Table 1 compares variational and MCMC estimates of the hyperparameters from a typical trial. For this trial, the variational solution closely matches Monte Carlo computations. The top row of Fig. 4 shows the posterior distribution of the hyperparameters produced by the variational (gray) and MCMC (black) methods in the same trial.

The main result of the first experiment is contained in Table 2. This table shows that the relative differences between variational and MCMC calculations are small, as predicted. These estimates closely correspond to the parameters $\log_{10}\sigma^2 = \log_{10} 9 \approx 0.95$, $\log_{10}\sigma_\beta^2 = \log_{10}(1/9) \approx -0.95$ and $\log_{10}\pi = \log_{10} 0.02 \approx -1.7$ used to simulate the data. It is still possible that closer agreement could be achieved by increasing the number of samples in the importance sampling part of the variational algorithm.

Since the variational method does not make assumptions about the posterior distribution of the hyperparameters, it is able to capture posterior correlations among the hyperparameters. For example, we expect that σ_β^2 and π are inversely correlated *a posteriori*; a smaller σ_β^2 corresponds to smaller effect sizes, which typically leads to more variables being included in the model, and a larger posterior estimate of π. Indeed, variational estimates of the posterior correlation coefficient of $\log_{10}\sigma_\beta^2$ and $\log_{10}\pi$ are -0.37 ± 0.14 over the 50 trials, and MCMC estimates for the same trials are -0.46 ± 0.16.

In addition to close agreement in point estimates of the hyperparameters, as Table 2 shows, posterior distributions also closely agree between the two inference methods. The bottom three rows of Fig. 4 show posterior distributions of the hyperparameters from trials that exhibit largest differences in the posterior mean estimates. Even in these worst cases, the variational approximation captures the correct overall shape and

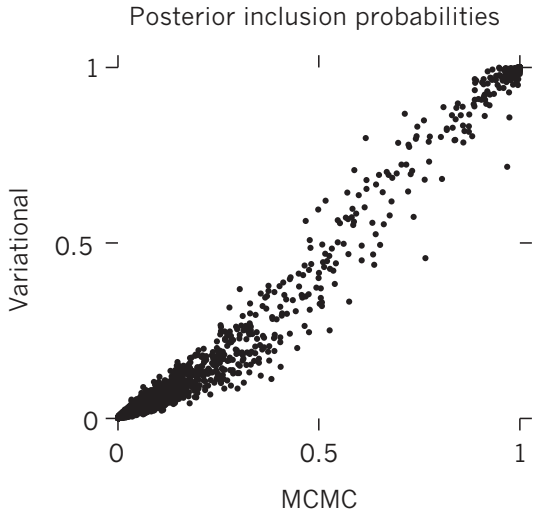

Figure 5: Scatter plot of posterior inclusion probabilities (PIPs) from the first simulation study. Each point is a posterior inclusion probability for one SNP in one trial. The horizontal axis is the MCMC estimate of the PIP, and the vertical axis is the variational estimate. Since there are 1000 SNPs in each simulation, and a total of 50 simulations, this plot has 50,000 points.

location, regardless of whether the posterior mass is diffuse or concentrated.

Not only do hyperparameter estimates agree, but Fig. 5 shows that the two methods also largely agree on posterior inclusion probabilities for the SNPs, particularly for the SNPs with high PIPs, which are the SNPs of greatest interest. If one were to select SNPs with PIPs at a certain threshold, the two methods would exhibit almost identical rates of false positives and false negatives (not shown).

Now that we've checked the accuracy of the variational method in the ideal setting when the variables are independent, next we investigate the accuracy of the variational method in the more realistic setting when many variables are strongly correlated.

5.2 "Targeted Region" study

Our second simulation study mimics a scenario in which a region of the genome has been identified from previous studies, and the goal is to identify genetic variants within this region that are relevant to the quantitative trait Y. The trait in this experiment is simulated, but we use actual samples of genetic variants, so this second experiment will better capture the patterns of correlations observed in genetic association studies. We assess the accuracy of the variational approximation in this setting.

For our simulations, we use SNPs from the ~10 megabase (Mb) region surrounding

	$E[\log_{10}\sigma^2]$	$E[\log_{10}\sigma_\beta^2]$	$E[\log_{10}\pi]$
variational	0.979	−0.928	−1.76
MCMC	0.972	−0.988	−1.69
difference	0.007	0.060	−0.07

Table 3: Posterior means of $\log_{10}$ hyperparameters for a typical trial in the "targeted region" simulation. See Table 1 for the legend.

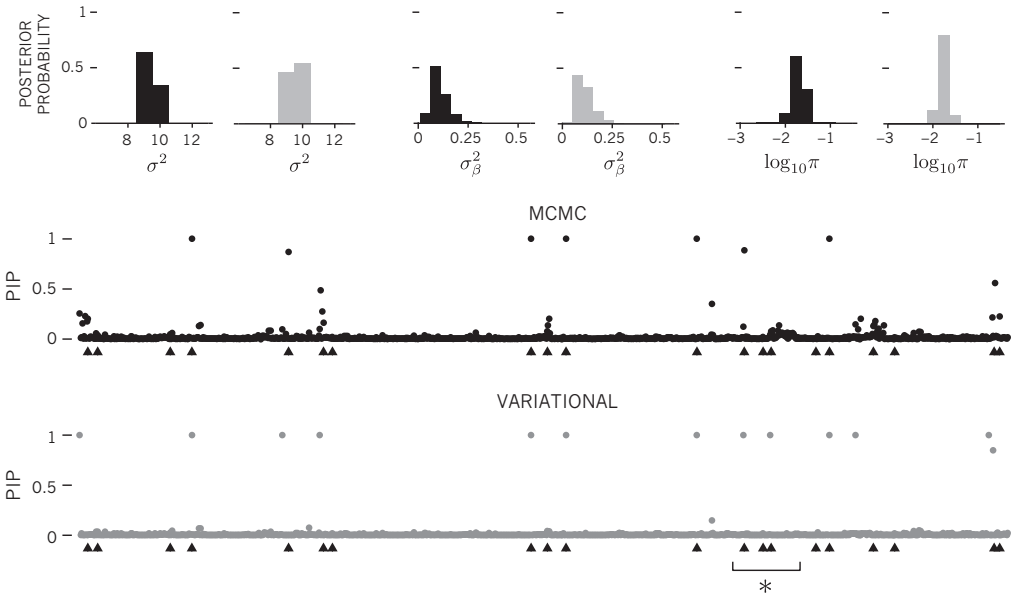

Figure 6: Results for a single trial, chosen to illustrate behaviour typical of the variational method in the "targeted region" simulation study. *Top panel:* posterior of hyperparameters. Variational estimates are in gray, MCMC computations are in black. *Middle and bottom panels:* Posterior inclusion probabilities (PIPs) for all SNPs in the targeted region. SNPs are ordered by their physical location on the chromosome. Black triangles mark the location of causal SNPs (SNPs that affect Y). The region marked with an asterisk ($*$) is shown in Fig. 7.

gene *IL27*. Genotypes of the 1037 SNPs lying in this region are taken from the cases and controls of the Wellcome Trust Case Control Consortium (2007) type 1 diabetes study. As before, we run 50 trials, with each data set of $n = 2000$ samples obtained by subsampling without replacement from the total of 4901 individuals. We use this data to simulate an artificial quantitative trait Y that is affected by 20 randomly-chosen SNPs, exactly as in the first simulation study. The variable selection model, priors, and implementation of the variational inference method remain unchanged from the first simulation study.

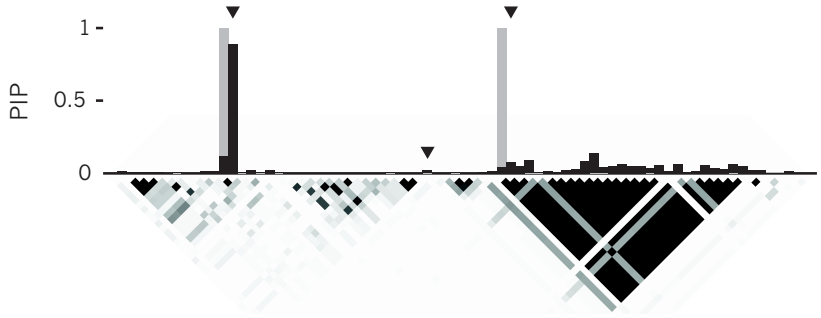

Figure 7: A closer look at the posterior inclusion probabilities in the region marked by the asterisk (∗) in Fig. 6. Variational and MCMC estimates are gray and black, respectively. Triangles mark the location of causal SNPs. Below the PIPs, the square of the correlation coefficient (r^2) is shown for every pair of SNPs; black indicates two SNPs are almost perfectly correlated, and white indicates no correlation.

First we examine results from a typical trial. Variational and MCMC estimates of the hyperparameters are given in Table 3, and estimates of the posterior distribution are shown in the top panel of Fig. 6. Remarkably, the accuracy of the variational approximation in this example is within range of the errors observed in the independent variables case; compare the differences reported in Table 3 to those in Table 2.

For the same trial, the middle and bottom rows of Fig. 6 show posterior inclusion probabilities (PIPs) for all SNPs in the targeted region, ordered by their location along the chromosome. Black triangles mark the locations of SNPs that affect Y (the "causal SNPs"). In this example, every SNP with a large variational PIP (bottom row) is inside a block of SNPs such that within this block there is a high probability, according to MCMC estimates, that at least one of the SNPs is included in the model. But within each of these blocks the variational approximation fails to capture uncertainty in the location of the selected SNP, akin to what we witnessed in Examples B and C in Sec. 4.

Consider, for example, the region indicated by the asterisk (refer to Figures 6 and 7). The SNP marked by the left-most black triangle in Fig. 7 is included in the model with high posterior probability (PIP = 0.89), whereas the variational approximation selects a neighbouring SNP with high probability (PIP = 1.00). The variational approximation has difficulty here with the strong correlation ($r = 0.95$) between the the two SNPs. On the right-hand side of Fig. 7, we are uncertain about the location of the causal variant because it is inside a block of highly correlated SNPs. As expected, the variational approximation fails to capture this uncertainty. But within this block of correlated SNPs, the variational approximation correctly calculates the *number* of SNPs included in the model; variational and MCMC estimates of the expected number of included SNPs are 1.08 and 1.19, respectively. Remember we are concerned with accuracy of computations, not accuracy of inferences, so the fact that variational approximation fails to select the causal SNP in each of these instances is not relevant.

	$E[\log_{10}\sigma^2]$	$E[\log_{10}\sigma_\beta^2]$	$E[\log_{10}\pi]$
variational	0.96 ± 0.03	-0.92 ± 0.29	-1.77 ± 0.17
MCMC	0.95 ± 0.03	-0.95 ± 0.29	-1.72 ± 0.17
mean diff.	0.007	0.027	-0.050
mean abs diff	0.009	0.053	0.056

Table 4: Posterior mean of $\log_{10}$ hyperparameters according to the variational and MCMC methods, averaged over 50 trials in the "targeted region" study. Standard error ($\pm$) is two times the sample deviation over the 50 trials. The top two rows show variational and MCMC estimates of the posterior expected values. The third and fourth rows show the mean of differences $\tilde{x} - x$, and the mean of absolute differences $|\tilde{x} - x|$, where x is the MCMC estimate and $\tilde{x}$ is the variational approximation.

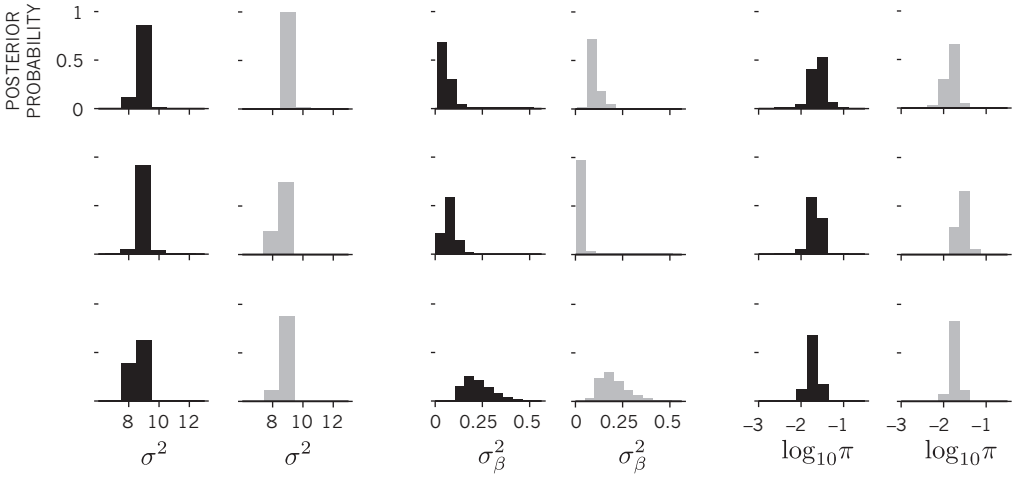

Figure 8: Posterior of $\log_{10}$ hyperparameters for several trials from the "targeted region" study. Variational estimates are gray, and MCMC computations are black. Each row corresponds to a single trial. Trials shown here were chosen because they exhibit the largest discrepancies between variational and MCMC estimates of the posterior mean of the hyperparameters.

Note that the posterior inclusion probabilities shown in Figures 6 and 7 correspond to one of several possible variational approximations; different starting points for the free parameters can produce slightly different answers, which correspond to different local minima of the Kullback-Leibler divergence.

In Table 4, we show results for all 50 simulations of the "targeted region" study. Variational estimates of $E[\log_{10}\sigma^2]$ and $E[\log_{10}\pi]$ appear to be reasonably accurate and, in fact, they are no worse than variational estimates in the setting with independent variables; compare these numbers with those in Table 2. This result makes sense in light of our discussion from Sec. 4, where we pointed out that the posterior of π will

be accurate so long as the variational approximation recovers the correct number of selected variables. Admittedly, the accuracy of the variational computations in this experiment may be attributed in part to an increase in the number of samples from $n = 500$ to $n = 2000$, as variational estimates tend to be more accurate when the posterior mass is concentrated. (If we kept n the same for this experiment, the posterior would be more diffuse because we have less information from correlated variables.) Fig. 8 shows posterior distributions of the hyperparameters from trials that showed the largest discrepancy between variational and MCMC estimates of the posterior means.

In summary, the main qualitative difference between the MCMC and variational inferences is that, when multiple correlated variables are associated with the outcome Y, the MCMC solution appropriately disperses the posterior probability across the correlated variables so that each one has a small PIP. In contrast, the variational approximation tends to concentrate the posterior probability onto a single variable, resulting in one large PIP, while the rest of the PIPs are near zero. This behaviour, which is also apparent in the work of Logsdon et al. (2010), can be viewed as a natural extension of the behaviour we observed in the toy examples (Sec. 4). In our simulations, MCMC estimates of PIPs tend to better reflect uncertainty in which variables should be included and, we presume, are closer to exact PIPs. Nonetheless, once one is aware of this feature of the variational approximation, the PIPs produced by the variational inference procedure can be useful because they correctly point to groups of correlated variables. For a genetic association study, this means that variational estimates of PIPs will single out the correct genomic region, if not the correct individual variant.

6 Case study: discovery of genome-wide associations for Crohn's disease

Now that we have assessed the accuracy of the variational approximation in simulations with independent and dependent variables, we illustrate its application to a large-scale variable selection problem with $\sim 400,000$ variables.

Genetic variants in genome-wide association studies are typically analyzed individually, ignoring correlations between variants. There are two reasons why it is beneficial to pursue a Bayesian hierarchical approach and analyze variants jointly. First, small genetic effects are sometimes easier to detect after accounting for factors that have a relatively strong effect on Y. Second, the conclusions of a genome-wide association study are influenced by our prior beliefs, and one way to improve objectivity is to infer hyperparameters from joint analysis of the data. Variational inference has the potential to realize the advantages of the Bayesian approach, and at a substantially reduced computational cost compared with MCMC inference. Here we compare analyses of a genome-wide association study using variational and MCMC inference approaches.

Our example is a case-control study of Crohn's disease, a common inflammatory bowel disease known to have a complex genetic basis. Recent analyses of genome-wide association studies have connected a large number of genetic variants to Crohn's

disease (Barrett et al. 2008; Franke et al. 2010). Although the variants identified so far account for only a portion of the variance in disease risk, many of these variants are believed to play an important biological role in signaling pathways that regulate responses to pathogens (Cho 2008). Our analysis is unlikely to offer new insights into Crohn's disease as findings have already been published based on the data we use here (WTCCC 2007). Nonetheless, these data provide a useful case study for illustrating the Bayesian hierarchical approach to analysis.

In this study, we have a total of $p = 442,001$ genetic variants (specifically, SNPs) on autosomal chromosomes. This is after applying quality control filters as described in WTCCC (2007), and after removing SNPs that exhibit no variation. We estimate any missing genotypes at these SNPs using the posterior mean minor allele count provided by BIMBAM (Servin and Stephens 2007), using SNP data from the International HapMap Consortium (2007).

The data from the genome-wide association study are the genotypes $\mathbf{X}$ and case-control labels y from a cohort of $n = 4686$ individuals. The 1748 subjects who carry the disease ("cases") are labeled $y_i = 1$, and the remaining 2938 disease-free subjects ("controls") are labeled $y_i = 0$. More details on this data can be found in WTCCC (2007).

To model case-control status, we replace the linear model for Y with a logistic regression. Under the logistic model, e^{β_k} is the "odds ratio" for locus k, the increase or decrease in disease odds for each copy of the minor allele. Implementation details of our variational method for the logistic model are given in the appendix. Otherwise, we conduct our analysis using the variable selection model as it is described in Sec. 2. Note that hyperparameter σ^2 is not needed for case-control data.

Next we discuss the choice of prior on the hyperparameters $\theta = \{\sigma_\beta^2, \pi\}$. Since this Crohn's disease study contains strong evidence for genetic risk factors, sensitivity of the final results to the prior on the hyperparameters is not a great concern here. But generally speaking it is important to choose this prior carefully because the data from a genetic association study may be only weakly informative.

Earlier, we expressed concern with the beta prior for π. Instead we adopt a normal prior on $\text{logit}_{10}\pi = \log_{10}(\frac{\pi}{1-\pi})$. We expect that only a small portion of the genetic factors increase (or decrease) susceptibility to Crohn's disease, so we set the prior mean to -5. This corresponds to 1 selected variable for every 100,000 SNPs, or a total of 4 or 5 causal variants. We set the prior standard deviation to 0.6, so that 0 to 70 causal variants are expected within the 95% prior credible interval.

We adopt a uniform prior on the proportion of variance explained, as we described in Sec. 5.1, except that we do not replace π by a constant in the expression for $\hat{\sigma}_z^2$. Therefore, σ_β^2 depends on π *a priori*.

We compute importance weights for $\hat{r}^2$ (the proportion of variance explained, as defined in Sec. 5.1) and $\text{logit}_{10}\pi$ at regular intervals of 0.05 and 0.25, respectively. Again, these intervals were chosen after some trial and error. We conduct importance sampling on $\hat{r}^2$ rather than σ_β^2 because it is easier to choose a reasonable range of values

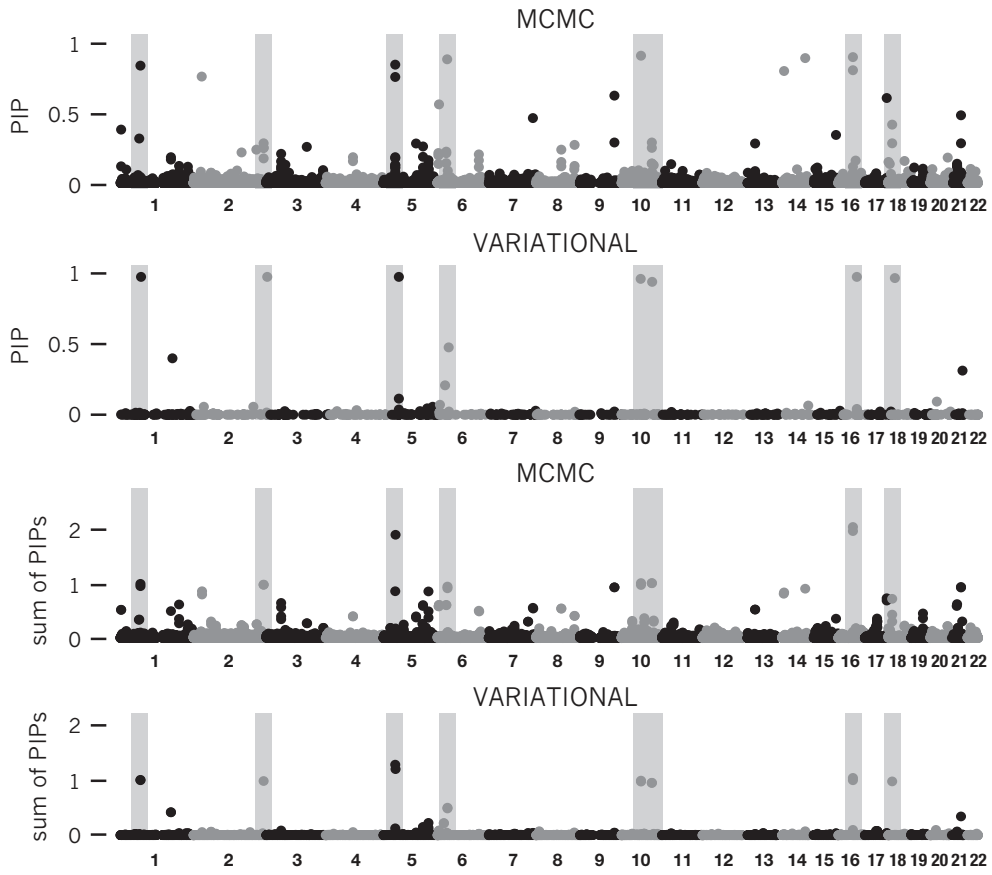

Figure 9: *Top two panels:* posterior inclusion probabilities for all SNPs in the genome-wide study of Crohn's disease. SNPs are ordered by chromosome, then by position along the chromosome. Autosomal chromosomes 1 through 22 are shown in alternating shades. Gray regions are the strongest associations identified in the original study (Table 3 in WTCCC 2007). *Bottom two panels:* sums of PIPs calculated over 200 kb segments.

for the proportion of variance explained.

We implemented our inference algorithm in MATLAB, and ran it on a machine with a 2.5 GHz Intel Xeon CPU. On average, coordinate ascent updates of the inner loop took 25 minutes to converge to a solution, though there was considerable variation in run time; depending on the choice of hyperparameters, the inner loop took as little as 9 minutes or as much as an hour to complete. It took about a day to complete the full variational inference procedure.

After running variational inference, we find that the posterior mean of σ_β is 0.201, with a posterior standard deviation of 0.05. The posterior mean of $\text{logit}_{10}\pi$ is -4.1, or

		sum of PIPs			
chr	pos. (Mb)	MCMC	Var.	SNP	PIP
1	67.3	1.001	1.015	rs11805303	1.000
2	233.8	0.985	1.001	rs10210302	1.000
5	40.3	2.598	1.301	rs17234657	1.000
6	32.7	0.950	0.508	rs9469220	0.489
9	114.4	0.937	0.045	rs4263839	0.026
10	64.0	1.016	1.008	rs10995271	0.984
10	101.1	1.015	0.965	rs7095491	0.963
14	96.4	0.911	0.071	rs11627513	0.068
16	49.3	2.050	1.013	rs17221417	1.000
18	12.7	0.723	0.991	rs2542151	0.990
21	39.2	0.940	0.345	rs2836753	0.321

Table 5: Regions of the genome with strong evidence of risk factors for Crohn's disease. Each row in the table is a 200 kb genomic segment for which the variational or MCMC estimate of the expected number of included SNPs ("sum of PIPs") exceeds 0.9. Rows highlighted in gray are the strongest associations identified in the original study (Table 3 in WTCCC 2007). Columns from left to right are: (1) chromosome number; (2) position of the start of the segment in megabases; (3) MCMC estimate of sum of SNPs; (4) variational estimate of sum of SNPs; (5) refSNP identifier for the SNP with the largest PIP in the segment, according to the variational method; (6) PIP of this SNP. All SNP information is based on human genome assembly 17 (NCBI build 35).

$\pi \approx 7/100,000$, with a posterior standard deviation of 0.2. This result suggests that, on average, about 30 SNPs are useful for predicting an individual's susceptibility to Crohn's disease, though the odds ratios e^{β_k} for many of these SNPs are close to one.

Ultimately, the aim of a genome-wide association study is to identify genetic variants and regions of the genome that affect disease outcome. For the remainder of our analysis, we focus on this aim. We compare the results from the variational method with findings from an analysis of the same data using the MCMC method described in Guan and Stephens (2011). (Results were kindly provided by Y. Guan; personal communication.) Considering the size of the variable selection problem, we should not assume that MCMC estimates are close to exact values.[4]

The top two panels in Fig. 9 show variational and MCMC estimates of the PIPs for all SNPs. From these two plots it is apparent that some PIPs coincide, but many do not; in other words, the two methods do not always agree on which SNPs might affect susceptibility to Crohn's disease. This is not surprising based our previous findings. As we discussed, when multiple correlated SNPs in a region are associated with Y, the variational approximation tends to select one of them and assign it a high PIP, whereas the MCMC approach divides the posterior probability among several correlated SNPs.

[4]MCMC with parallel tempering would yield more accurate inferences (Bottolo and Richardson 2010), but this would increase the already high computational cost for this problem.

Therefore, for a better comparison of the variational and MCMC methods, we ask whether the methods identify the same regions of the genome instead of the same SNPs. We divide the genome into 200 kilobase (kb) segments, in which each pair of neighbouring segments overlaps by 100 kb. On average, a 200 kb segment contains 37 SNPs. For each segment, we compute the the sum of the posterior inclusion probabilities or, equivalently, the expected number of SNPs associated with disease risk.

The bottom two panels in Fig. 9 show sums of PIPs across the genome. Table 5 lists all regions of the genome for which at least one of the two methods declares that the region contains a risk factor for Crohn's disease with high probability (the sum of PIPs exceeds 0.9). This table does not show overlapping segments that share the same association signal. As expected, Table 5 recapitulates the strongest associations with Crohn's disease identified in the original individual-SNP analysis—specifically, it recovers SNPs with trend p-values less than 4×10^{-8} in Table 3 of WTCCC (2007). Two SNPs from the original analysis showing slightly weaker associations in region 49.3–49.87 Mb on chromosome 3 and region 150.15–150.31 Mb on chromosome 5 do not satisfy our criterion for significance.

On the whole, the regions identified by the variational and MCMC methods in Table 5 coincide. But there are notable discrepancies. Three regions on chromosomes 9, 14 and 21 have high expected counts in the MCMC inference, but low counts according to the variational approximation. Interestingly, none of these three regions have shown up in large meta-analyses of Crohn's disease (Franke et al. 2010; Mathew 2008), suggesting that these may be false associations. Perhaps this is due to MCMC convergence issues. In contrast, the 12.7–12.9 Mb region on chromosome 18 that has a higher sum of PIPs under variational inference has been confirmed by the same meta-analyses. This latter region is a compelling candidate for Crohn's disease because it contains a gene for a T cell protein that plays a role in regulation of inflammatory responses to pathogens (Mathew 2008). While these results suggest that the regions identified by the variational method are more reliable than those identified by MCMC, we caution that this comparison is limited. For example, the two regions on chromosomes 3 and 5 that were identified in the original analysis (WTCCC 2007) and not listed in Table 5 are assigned higher expected counts by MCMC than by the variational method. These two regions were also confirmed by larger follow-up studies. Nonetheless, these results suggest that variational inference can be a useful and less costly alternative to MCMC in large variable selection problems.

7 Discussion

The main goal of this paper was to assess the utility of a variational approximation for Bayesian variable selection in large-scale problems. It is important to investigate alternatives to the standard approach—Markov chain Monte Carlo—to fitting variable selection models because MCMC is often difficult to implement effectively. Designing a Markov chain that efficiently explores the posterior distribution has been the focus of dozens of research articles over the past couple decades.

Our results highlight the pros and cons of the variational approach. A key advantage is its computational complexity, which is linear in the number of variables. (Actual run times depend on the number of coordinate ascent iterations needed to reach convergence, which can vary depending on context; ideal conditions for quick convergence are a sparse model and weakly correlated variables.) The variational method generally provides accurate posterior distributions for hyperparameters. In idealized situations with independent explanatory variables, it also provides accurate estimates of posterior inclusion probabilities. When variables are correlated, individual posterior inclusion probabilities are often inaccurate. Still, variational inferences can be useful in this case because they help identify relevant variables and, for genetic association studies, they point to relevant regions of the genome. And while this is not an aspect we have touched on in this manuscript, our results suggest that the variational approximation can be useful for prediction, particularly when we are less interested in identifying which variables are included in the predictive model of Y.

Building on Logsdon et al. (2010), the variational method we describe is very flexible. For example, it allows arbitrary priors for the hyperparameters, and continuous or binary outcomes (see the appendix). The ability to handle binary outcomes is particularly useful in genetic association studies, where case-control studies are common. This is a case where inference solutions based on MCMC can struggle: although data augmentation (Albert and Chib 1993) is a well-known strategy for coping with binary outcomes in MCMC, it often yields a slowly converging Markov chain (Liu and Wu 1999). Slow convergence is usually tolerated in small variable selection problems, but it can be a crippling issue for problems with thousands of variables.

We derived the variational approximation with a specific prior for β and γ, but it is easy to extend the approximation to other priors, including the g-prior (Liang et al. 2008; Zellner 1986). The variational approximation is appropriate for the g-prior without modification because β_j and β_k will be nearly independent *a posteriori* under the same conditions as before, when X_j and X_k are independent. It is possible that variational inference could be useful for other approaches to Bayesian variable selection, such as those based on normal-gamma priors (Griffin and Brown 2010), but this remains an open question.

To compute the posterior distribution of the hyperparameters without imposing additional variational approximations, we suggested using importance sampling in which the marginal likelihood in the importance weight is replaced with its corresponding best variational lower bound. This idea of using a variational bound to approximate the shape of the marginal likelihood has recently gained traction as a way to improve variational inference (Bouchard and Zoeter 2009; Cseke and Heskes 2011; Ormerod 2011) and, in principle, it could be useful for a wide variety problems. But in practice the accuracy of the variational bound needs to be assessed.

Importance sampling worked well for our applications because we had at most three hyperparameters. For a variable selection model with a large number of hyperparameters, other Monte Carlo strategies would probably be more effective. For example, one could replace the likelihood terms that appear in the Metropolis-Hastings accep-

tance probability (Chib and Greenberg 1995) with the corresponding variational lower bound. But this may lead to an expensive Metropolis-Hastings step, because computing the likelihood would involve running the coordinate ascent updates to completion. It remains to be seen whether this inference approach is useful for problems with many hyperparameters.

A natural extension to our work would be to develop approximations with less stringent conditional independence assumptions. This would be especially useful when we have prior knowledge about the conditional independence structure of the variables. For example, in genome-wide association studies the most strongly correlated SNPs are closest to each other on the chromosome. Nevertheless, the fully-factorized approximation we investigated in this paper remains appealing for its simplicity and ease of use.

Software

MATLAB and R implementations of our variational inference algorithm are available on the Stephens lab website.

Appendix: extension to case-control studies

In this section, we describe an extension to our variational inference method for problems with a binary outcome $Y \in \{0, 1\}$.

We begin with a linear model for the log-odds:

$$\log\left\{\frac{p(Y=1)}{p(Y=0)}\right\} = \beta_0 + \sum_{k=1}^{p} X_k\beta_k. \tag{15}$$

From this identity, it follows that binary outcome Y is a coin toss with success rate $\psi(\beta_0 + X^T\beta)$, where $\psi(x) = 1/(1+e^{-x})$ is the sigmoid function. Assuming independence of the samples y_i, and defining $p_i = \psi(\beta_0 + x_i^T\beta)$ to be the success rate for the ith sample, the likelihood is the product

$$p(y \mid \mathbf{X}, \beta_0, \beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}. \tag{16}$$

The scale parameter σ^2 is not needed for modeling a binary outcome, so we only have two hyperparameters (σ_β^2, π) for the variable selection model.

From a computation point of view, the main inconvenience of the logistic model is the appearance of the nonlinear sigmoid terms in the likelihood $p(y \mid \mathbf{X}, \beta_0, \beta)$. This will make it difficult to integrate over β_0 and β. Laplace's method is commonly used to approximate the integral by forming a Taylor series expansion to the logarithm of the posterior density function. This often results in a good approximation when the Taylor series expansion is centered about a mode of the posterior (Tierney and Kadane 1986). For variable selection, however, it is extraordinarily difficult—and probably not helpful—to compute a posterior mode due to the discontinuous spike and slab prior.

For MCMC inference, data augmentation (Albert and Chib 1993) is a natural way to deal with the nonlinear likelihood (this trick is typically used for probit regression). By cleverly introducing an auxiliary variable, the posterior of β becomes normal conditioned on that variable. But for variational inference it is unclear whether this auxiliary variable is helpful. Instead, we formulate an additional variational lower bound.

Skipping the derivation (see Bishop 2006 or Jaakkola and Jordan 2000 for details), the lower bound on the logarithm of the sigmoid function is

$$\log \psi(x) \geq \log \psi(\eta) + \tfrac{1}{2}(x - \eta) - \tfrac{u}{2}(x^2 - \eta^2), \tag{17}$$

where we have defined $u = \frac{1}{\eta}(\psi(\eta) - \frac{1}{2})$. This identity holds for any choice of η. Since this bound is symmetric about $\eta = 0$, we restrict η to the non-negative numbers.

For the moment, assume that the intercept β_0 is zero. Replacing the sigmoid terms in the likelihood (16) by their lower bound, we obtain a bound on the marginal likelihood for a given collection of free parameters $\eta = (\eta_1, \ldots, \eta_n)$:

$$\begin{aligned} p(y \,|\, \mathbf{X}, \theta, \beta_0 = 0) &= \iint p(y \,|\, \mathbf{X}, \beta_0 = 0, \beta)\, p(\beta, \gamma \,|\, \theta)\, d\beta\, d\gamma \\ &\geq \iint e^{f(\beta;\eta)}\, p(\beta, \gamma \,|\, \theta)\, d\beta\, d\gamma, \end{aligned} \tag{18}$$

where we define

$$f(\beta; \eta) \equiv \sum_{i=1}^{n} \log \psi(\eta_i) + \tfrac{\eta_i}{2}(u_i \eta_i - 1) - \tfrac{1}{2}\beta^T \mathbf{X}^T U \mathbf{X} \beta + (y - \tfrac{1}{2})^T \mathbf{X} \beta, \tag{19}$$

and where U is the $n \times n$ matrix with diagonal entries u_i. Notice that (19) is a quadratic function of β. If the prior on β were, say, normal with zero mean and covariance Σ_0, then the variational approximation to the posterior would be normal with mean $\mu = \Sigma \mathbf{X}^T (y - \frac{1}{2})$ and covariance $\Sigma = (\Sigma_0^{-1} + \mathbf{X}^T U \mathbf{X})^{-1}$, and we would have an analytic expression for the lower bound (18).

This variational approximation is similar to Laplace's method in the sense that it reweights the rows of $\mathbf{X}$ by scalars u_i. Another interesting outcome from the variational approximation is that $y - \frac{1}{2}$ acts as a vector of continuous observations.

A natural question at this point is how to adjust the free parameters $\eta = (\eta_1, \ldots, \eta_n)$ so that the lower bound (18) is as tight as possible. We express the solution using expectation maximization (EM): in the E-step, compute expectations of the unknowns (β, γ), which we do in an approximate manner using the variational method; in the M-step, compute the value of η that maximizes the expected value of the log-density $f(\beta; \eta)$. Note that while we formulate the variational inference algorithm using EM, we are not using EM in the conventional sense. The argument we are maximizing over, η, is not a parameter of the model; it is only a vector parameterizing the variational approximation and does not have a meaningful interpretation beyond that.

To derive the M-step, we take partial derivatives of $E[f(\beta; \eta)]$ with respect to the free parameters:

$$\frac{\partial E[f(\beta; \eta)]}{\partial \eta_i} = \tfrac{1}{2}(\eta_i^2 - (x_i^T \mu)^2 - x_i^T \Sigma x_i) \times \frac{du_i}{d\eta_i}, \tag{20}$$

where x_i is the ith row of $\mathbf{X}$, and μ and Σ are the posterior mean and covariance of β (which we computed in the E-step). We can ignore the prior $p(\beta, \gamma)$ in the M-step because it is unaffected by the choice of η. Taking note that u_i is a strictly monotonic function of η_i, the fixed point and M-step update for η_i is

$$\eta_i^2 = (x_i^T \mu)^2 + x_i^T \Sigma x_i. \tag{21}$$

This expression will simplify once we apply the variational approximation.

For linear regression, we remove the effect of the intercept β_0 by centering y and the columns of $\mathbf{X}$ so that they each have a mean of zero. Next we explain how to accomplish this for the variational approximation to logistic regression. This can be understood as a generalization of centering $\mathbf{X}$ and y with weighted samples.

Suppose the prior on β_0 is normal with zero mean and standard deviation σ_0. At the limit as σ_0 becomes large (yielding an improper prior on β_0), the lower bound to the marginal likelihood times σ_0 is

$$\begin{aligned} \sigma_0\, p(y \mid \mathbf{X}, \theta) &= \sigma_0 \iiint p(y \mid \mathbf{X}, \beta_0, \beta)\, p(\beta, \gamma \mid \theta)\, p(\beta_0)\, d\beta_0\, d\beta\, d\gamma \\ &\geq \hat{\sigma}_0 \iint e^{\hat{f}(\beta;\eta)}\, p(\beta, \gamma \mid \theta)\, d\beta\, d\gamma, \end{aligned} \tag{22}$$

where we define

$$\hat{f}(\beta; \eta) \equiv \sum_{i=1}^{n} \log \psi(\eta_i) + \tfrac{\eta_i}{2}(u_i \eta_i - 1) - \tfrac{1}{2}\beta^T \mathbf{X}^T \hat{U} \mathbf{X} \beta + \hat{y}^T \mathbf{X} \beta + \tfrac{1}{2}\bar{y}^2/\bar{u}, \tag{23}$$

and $\hat{\sigma}_0 = 1/\sqrt{\bar{u}}$ is the standard deviation of the intercept β_0 given β. We write the posterior mode of the intercept when $\beta = 0$ as $\hat{\beta}_0 = \bar{y}/\bar{u}$, and we define

$$\begin{aligned} \hat{U} &= U - \tfrac{uu^T}{\bar{u}} & \hat{y} &= y - \tfrac{1}{2} - \hat{\beta}_0 u \\ \bar{u} &= \textstyle\sum_{i=1}^n u_i & \bar{y} &= \textstyle\sum_{i=1}^n (y_i - \tfrac{1}{2}). \end{aligned}$$

Notice that the entries of vector $\hat{y}$ sum to zero regardless of the value of u. Also note that when $U = I$, replacing y with $\hat{y}$ and $\mathbf{X}^T U \mathbf{X}$ with $\mathbf{X}^T \hat{U} \mathbf{X}$ is equivalent to centering y and the columns of $\mathbf{X}$.

Up to this point, we have yet to incorporate the fully-factorized variational approximation $q(\beta, \gamma)$ into our inference procedure for the logistic regression model. Since we've taken care to integrate out the intercept from the variational lower bound, it can be shown that the expected value of any off-diagonal entry (j, k) of $\mathbf{X}^T \hat{U} \mathbf{X}$ is zero whenever variables X_j and X_k are conditionally independent. Like we did for the variable selection model in linear regression, we can apply the fully-factorized approximation (6) to the integral (22), yielding an additional lower bound. Proceeding in a similar

manner, we obtain the following analytical expression for the variational lower bound:

$$\begin{aligned}\log p(y \mid \mathbf{X}, \theta) \geq \log \hat{\sigma}_0 + \tfrac{1}{2}\bar{y}^2/\bar{u} + \sum_{i=1}^{n} \log \psi(\eta_i) + \tfrac{\eta_i}{2}(u_i\eta_i - 1) + \hat{y}^T\mathbf{X}r - \tfrac{1}{2}r^T\mathbf{X}^T\hat{U}\mathbf{X}r \\ - \frac{1}{2}\sum_{k=1}^{p}(\mathbf{X}^T\hat{U}\mathbf{X})_{kk}\mathbb{V}\mathrm{ar}[\beta_k] + \sum_{k=1}^{p}\frac{\alpha_k}{2}\left[1 + \log\left(\frac{s_k^2}{\sigma_\beta^2}\right) - \frac{s_k^2 + \mu_k^2}{\sigma_\beta^2}\right] \\ - \sum_{k=1}^{p}\alpha_k \log\left(\frac{\alpha_k}{\pi}\right) - \sum_{k=1}^{p}(1-\alpha_k)\log\left(\frac{1-\alpha_k}{1-\pi}\right), \qquad (24)\end{aligned}$$

and the approximate solution is given by coordinate ascent equations

$$\mathbb{V}\mathrm{ar}[\beta_k \mid \gamma_k = 1] \approx s_k^2 = \frac{1}{(\mathbf{X}^T\hat{U}\mathbf{X})_{kk} + 1/\sigma_\beta^2} \qquad (25)$$

$$E[\beta_k \mid \gamma_k = 1] \approx \mu_k = s_k^2\left((\mathbf{X}^T\hat{y})_k - \sum_{j\neq k}(\mathbf{X}^T\hat{U}\mathbf{X})_{jk}\alpha_j\mu_j\right) \qquad (26)$$

$$\frac{p(\gamma_k = 1 \mid \mathbf{X}, y, \theta)}{p(\gamma_k = 0 \mid \mathbf{X}, y, \theta)} \approx \frac{\alpha_k}{1-\alpha_k} = \frac{\pi}{1-\pi} \times \frac{s_k}{\sigma_\beta} \times e^{\mathrm{SSR}_k/2}. \qquad (27)$$

These are the coordinate descent updates that minimize the Kullback-Leibler divergence for the fully-factorized approximation to the logistic model, in which we place spike and slab priors on the regression coefficients.

Under the fully-factorized variational approximation, with the intercept included in the logistic regression, the M-step update for the free parameters η, from (21), becomes

$$\eta_i^2 = \left(E[\beta_0] + \textstyle\sum_k x_{ik}E[\beta_k]\right)^2 + \mathbb{V}\mathrm{ar}[\beta_0] + \displaystyle\sum_{k=1}^{p} x_{ik}^2\mathbb{V}\mathrm{ar}[\beta_k] + 2\sum_{k=1}^{p} x_{ik}\mathbb{C}\mathrm{ov}[\beta_0, \beta_k]. \qquad (28)$$

Means and variances of the coefficients β are easily obtained from the variational approximation: $E[\beta_k] = \alpha_k\mu_k$ and $\mathbb{V}\mathrm{ar}[\beta_k] = \alpha_k(s_k^2 + \mu_k^2) - (\alpha_k\mu_k)^2$. The remaining means and covariances in the above expression are

$$E[\beta_0] = \hat{\sigma}_0^2(\bar{y} - u^T X E[\beta]) \qquad (29)$$

$$\mathbb{V}\mathrm{ar}[\beta_0] = \hat{\sigma}_0^2\left(1 + \hat{\sigma}_0^2\textstyle\sum_k (X^T u)_k^2\mathbb{V}\mathrm{ar}[\beta_k]\right) \qquad (30)$$

$$\mathbb{C}\mathrm{ov}[\beta_0, \beta_k] = -\hat{\sigma}_0^2(X^T u)_k\mathbb{V}\mathrm{ar}[\beta_k]. \qquad (31)$$

Appendix: details of Markov chain Monte Carlo method

To simulate the Markov chain, we first analytically integrate out the additive effects β and prior π:

$$\begin{aligned}p(\gamma, \sigma^2, \sigma_\beta^2 \mid \mathbf{X}, y) &\propto p(y \mid \mathbf{X}, \gamma, \sigma^2, \sigma_\beta^2)\, p(\gamma, \sigma^2, \sigma_\beta^2) \\ &= \iint p(y \mid \mathbf{X}, \beta, \sigma^2)\, p(\beta \mid \gamma, \sigma^2, \sigma_\beta^2)\, p(\gamma \mid \pi)\, p(\pi)\, p(\sigma^2, \sigma_\beta^2)\, d\beta\, d\pi. \qquad (32)\end{aligned}$$

In the development of the MCMC algorithm, we assume that the prior on π is beta with prior sample sizes a and b.

Each iteration of the Markov chain consists of three Metropolis-Hastings steps within a Gibbs sampler (Chib and Greenberg 1995): (1) adjust σ^2 given γ and σ_β^2; (2) adjust σ_β^2 given γ and σ^2; and (3) adjust γ given σ^2 and σ_β^2. Assuming the prior on σ^2 is inverse gamma with shape $a_\sigma/2$ and scale $b_\sigma/2$—the standard prior $p(\sigma^2) \propto 1/\sigma^2$ is the limiting density of the inverse gamma as a_σ and b_σ approach zero—we have a Gibbs sampling step for σ^2 because $\sigma^2 \,|\, \gamma, \sigma_\beta^2$ is inverse gamma with shape $(a_\sigma + n)/2$ and scale $(b_\sigma + y^T y - \mathrm{SSR})/2$, in which $\mathrm{SSR} = y^T \mathbf{X} S \mathbf{X}^T y$ is the sum of squares due to regression, and $S = (1/\sigma_\beta^2 I + \mathbf{X}^T\mathbf{X})^{-1}$ times σ^2 is the covariance matrix of the nonzero coefficients. (Note that the mean of the nonzero coefficients is $S\mathbf{X}^T y$.) The $\mathbf{X}$ in these expressions only includes the columns corresponding to nonzero coefficients.

To update σ_β^2 given γ and σ^2, we propose a new candidate $\hat{\sigma}_\beta^2 = \sigma_\beta^2 e^u$, where u is a random draw from the standard normal, and admit this candidate into the Markov chain with Metropolis-Hastings acceptance probability

$$\mathcal{A}(\sigma_\beta^2, \hat{\sigma}_\beta^2) = \min\left\{1, \frac{p(\hat{\sigma}_\beta^2, \sigma^2)}{p(\sigma_\beta^2, \sigma^2)} \times \exp\left(\frac{\hat{\mathrm{SSR}} - \mathrm{SSR}}{2\sigma^2}\right) \times \left|\frac{\hat{S}}{S}\right|^{1/2} \times \frac{\hat{\sigma}_\beta^2}{\sigma_\beta^2}\right\}. \tag{33}$$

Since the samples σ_β^2 and $\hat{\sigma}_\beta^2$ depend on each other, the extra term $\hat{\sigma}_\beta^2/\sigma_\beta^2$ is needed so that the Metropolis-Hastings acceptance probability satisfies the detailed balance condition, as we need to account for the change of variables (Green 2003).

The most complicated part of our algorithm is the Metropolis-Hastings step for γ given σ^2 and σ_β^2. The proposal is as follows. First, we decide whether to add a variable (a "birth") or remove a variable (a "death") from the model according to probabilities $q_{\mathsf{birth}}(\gamma)$ and $q_{\mathsf{death}}(\gamma)$, respectively, such that $q_{\mathsf{birth}}(\gamma) + q_{\mathsf{death}}(\gamma) = 1$. When the number of variables included in the model ($m = |\gamma|$) is greater than zero and less than the total number of variables (p), we conduct a birth move or a death move with equal probability. When $m = 0$, $q_{\mathsf{birth}}(\gamma) = 1$. And when $m = p$, $q_{\mathsf{death}}(\gamma) = 1$.

First consider the case when we have chosen to add a variable to the model. Instead of selecting a variable at random from the pool of excluded variables (variables X_k for which $\gamma_k = 0$), we conduct a more efficient move and select a variable at a frequency proportional to the likelihood. Precisely, we propose that $\gamma_k = 0$ be switched to $\hat{\gamma}_k = 1$ with probability proportional to $p(y \,|\, \mathbf{X}, \gamma_k = 1, \gamma_{-k}, \sigma^2, \sigma_\beta^2)$. Writing $\ell(\gamma)$ as shorthand for the likelihood,

$$\ell(\gamma) \equiv p(y \,|\, \mathbf{X}, \gamma, \sigma^2, \sigma_\beta^2) = (2\pi\sigma^2)^{-n/2} |S/\sigma_\beta^2|^{1/2} \exp\left\{\tfrac{1}{2\sigma^2}(\mathrm{SSR} - y^T y)\right\}, \tag{34}$$

the Metropolis-Hastings acceptance probability for the birth move works out to be

$$\mathcal{A}(\gamma_k=0,\hat{\gamma}_k=1) = \min\left\{1, \frac{q_{\mathsf{death}}(\hat{\gamma})}{q_{\mathsf{birth}}(\gamma)} \times \frac{a+m}{b+p-m-1} \times \frac{\ell(\gamma_k=0,\gamma_{-k})}{\ell(\gamma_k=1,\gamma_{-k})} \times \sum_{j:\gamma_j=0} \frac{\ell(\gamma_j=1,\gamma_{-j})}{\ell(\gamma_j=0,\gamma_{-j})} \Big/ \sum_{j:\hat{\gamma}_j=1} \frac{\ell(\gamma_j=0,\hat{\gamma}_{-j})}{\ell(\gamma_j=1,\hat{\gamma}_{-j})}\right\}, \tag{35}$$

where γ_{-k} is the set of all indicator variables except the kth one. For the death move, we propose that a $\gamma_k = 1$ be flipped to $\hat{\gamma}_k = 0$ with probability proportional to $\ell(\gamma_k = 0, \gamma_{-k})$. The acceptance probability for a death move is

$$\mathcal{A}(\gamma_k=1,\hat{\gamma}_k=0) = \min\left\{1, \frac{q_{\mathsf{birth}}(\hat{\gamma})}{q_{\mathsf{death}}(\gamma)} \times \frac{b+p-m-2}{a+m-1} \times \frac{\ell(\gamma_k=1,\gamma_{-k})}{\ell(\gamma_k=0,\gamma_{-k})} \times \sum_{j:\gamma_j=1} \frac{\ell(\gamma_j=0,\gamma_{-j})}{\ell(\gamma_j=1,\gamma_{-j})} \Big/ \sum_{j:\hat{\gamma}_j=0} \frac{\ell(\gamma_j=1,\hat{\gamma}_{-j})}{\ell(\gamma_j=0,\hat{\gamma}_{-j})}\right\}. \tag{36}$$

Computing the Metropolis-Hastings acceptance probabilities can be done efficiently with a judicious block decomposition of the determinant and inverse of matrix S.

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. 98, 100

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). "An introduction to MCMC for machine learning." *Machine Learning*, 50: 5–43. 74

Attias, H. (1999). "Independent factor analysis." *Neural Computation*, 11(4): 803–851. 75, 79

Barrett, J. C., Hansoul, S., Nicolae, D. L., et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." *Nature Genetics*, 40(8): 955–962. 94

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag, 2nd edition. 86

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 100

Blei, D., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. 75, 81

Bottolo, L. and Richardson, S. (2010). "Evolutionary stochastic search for Bayesian model exploration." *Bayesian Analysis*, 5: 583–618. 74, 77, 96

Bouchard, G. and Zoeter, O. (2009). "Split variational inference." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 57–64. 75, 98

Chib, S. and Greenberg, E. (1995). "Understanding the Metropolis-Hastings algorithm." *American Statistician*, 49(4): 327–335. 99, 103

Chipman, H., George, E. I., and McCulloch, R. E. (2001). "The practical implementation of Bayesian model selection." In *Model Selection*, volume 38 of *IMS Lecture Notes*, 65–116. 76

Cho, J. H. (2008). "The genetics and immunopathogenesis of inflammatory bowel disease." *Nature Reviews Immunology*, 8(6): 458–466. 94

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 19(1): 81–94. 76

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational and Graphical Statistics*, 20(1): 80–101. 74, 77

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, 2nd edition. 78

Cseke, B. and Heskes, T. (2011). "Approximate marginals in latent Gaussian models." *Journal of Machine Learning Research*, 12: 417–454. 75, 81, 98

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). "On Bayesian model and variable selection using MCMC." *Statistics and Computing*, 12: 27–36. 74, 77

Franke, A., McGovern, D. P. B., Barrett, J. C., et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." *Nature Genetics*, 42(12): 1118–1125. 94, 97

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 74

— (1997). "Approaches to Bayesian variable selection." *Statistica Sinica*, 7: 339–373. 73

Green, P. J. (2003). "Trans-dimensional Markov chain Monte Carlo." In *Highly Structured Stochastic Systems*. Oxford University Press. 103

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5: 171–188. 76, 98

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(3): 1780–1815. 74, 86, 96

He, Q. and Lin, D. (2011). "A variable selection method for genome-wide association studies." *Bioinformatics*, 27(1): 1–8. 74

Heskes, T., Zoeter, O., and Wiegerinck, W. (2004). "Approximate expectation maximization." In *Advances in Neural Information Processing Systems*, volume 16, 353–360. 75

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). "Simultaneous

analysis of all SNPs in genome-wide and re-sequencing association studies." *PLoS Genetics*, 4(7): e1000130. 74

International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." *Nature*, 449(7164): 851–861. 94

Jaakkola, T. S. and Jordan, M. I. (2000). "Bayesian parameter estimation via variational methods." *Statistics and Computing*, 10: 25–37. 100

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine Learning*, 37: 183–233. 73, 74, 81

Khan, M. E., Marlin, B., Bouchard, G., and Murphy, K. (2010). "Variational bounds for mixed-data factor analysis." In *Advances in Neural Information Processing Systems 23*, 1108–1116. 81

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. 77, 98

Liu, J. S. and Wu, Y. N. (1999). "Parameter expansion for data augmentation." *Journal of the American Statistical Association*, 94(448): 1264–1274. 98

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010). "A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis." *BMC Bioinformatics*, 11(1): 58. 74, 75, 79, 93, 98

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. 83

Mathew, C. G. (2008). "New links to the pathogenesis of Crohn disease provided by genome-wide association scans." *Nature Reviews Genetics*, 9(1): 9–14. 97

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). "p-values for high-dimensional regression." *Journal of the American Statistical Association*, 104(488): 1671–1681. 74

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83: 1023–1032. 77

O'Hara, R. B. and Sillanpää, M. J. (2009). "A review of Bayesian variable selection methods: what, how and which." *Bayesian Analysis*, 4: 85–118. 76

Ormerod, J. T. (2011). "Grid based variational approximations." *Computational Statistics and Data Analysis*, 55(1): 45–56. 75, 81, 98

Ormerod, J. T. and Wand, M. P. (2010). "Explaining variational approximations." *The American Statistician*, 64(2): 140–153. 74

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 76

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for

linear regression models." *Journal of the American Statistical Association*, 92(437): 179–191. 73

Servin, B. and Stephens, M. (2007). "Imputation-based analysis of association studies: candidate regions and quantitative traits." *PLoS Genetics*, 3(7): e114. 74, 77, 94

Stephens, M. and Balding, D. J. (2009). "Bayesian statistical methods for genetic association studies." *Nature Reviews Genetics*, 10(10): 681–690. 74

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *Journal of the Royal Statistical Society: Series B*, 58(1): 267–288. 74

— (2011). "Regression shrinkage and selection via the Lasso: a retrospective." *Journal of the Royal Statistical Society: Series B*, 73(3): 273–282. 74

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393): 82–86. 99

Turner, R., Berkes, P., and Sahani, M. (2008). "Two problems with variational expectation maximisation for time-series models." In Barber, D., Cemgil, A. T., and Chiappa, S. (eds.), *Proceedings of the Inference and Estimation in Probabilistic Time-Series Models Workshop*, 107–115. 83

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical models, exponential families, and variational inference." *Foundations and Trends in Machine Learning*, 1: 1–305. 74

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature*, 447: 661–678. 90

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). "Genome-wide association analysis by Lasso penalized logistic regression." *Bioinformatics*, 25(6): 714–721. 74

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Goal, P. K. and Zellner, A. (eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. Edward Elgar Pub. Ltd. 77, 98

Acknowledgments

We thank an anonymous referee and the associate editor for their helpful comments on the original manuscript. We thank Firas Hamze, Bryan Howie, Xiaoquan Wen and Xiang Zhou for helpful discussions, Kevin Bullaughey and John Zekos for expert technical support, and Yongtao Guan for assistance with the Crohn's disease case study. This work was supported by a grant from the National Institute of Health (HG02585), and a postdoctoral fellowship from the Human Frontiers Science Program.