Estimation of Quantitative Trait Locus Effects with Epistasis by Variational Bayes Algorithms

Zitong Li* and Mikko J. Sillanpää*

Departments of *Mathematics and Statistics and §Agricultural Sciences, University of Helsinki, Helsinki FIN-00014, Finland, and Departments of ‡Mathematical Sciences and †Biology, University of Oulu, Oulu FIN-90014, Finland

ABSTRACT Bayesian hierarchical shrinkage methods have been widely used for quantitative trait locus (QTL) mapping. From the computational perspective, the application of the Markov chain Monte Carlo (MCMC) method is not optimal for high-dimensional problems such as the ones arising in epistatic analysis. Maximum a posteriori (MAP) estimation can be a faster alternative, but it usually produces only point estimates without providing any measures of uncertainty (i.e., interval estimates). The variational Bayes method, stemming from the mean field theory in theoretical physics, is regarded as a compromise between MAP and MCMC estimation, which can be efficiently computed and produces the uncertainty measures of the estimates. Furthermore, variational Bayes methods can be regarded as the extension of traditional expectation-maximization (EM) algorithms and can be applied to a broader class of Bayesian models. Thus, the use of variational Bayes algorithms based on three hierarchical shrinkage models including Bayesian adaptive shrinkage, Bayesian LASSO, and extended Bayesian LASSO is proposed here. These methods performed generally well and were found to be highly competitive with their MCMC counterparts in our example analyses. The use of posterior credible intervals and permutation tests are considered for decision making between quantitative trait loci (QTL) and non-QTL. The performance of the presented models is also compared with R/qtlbim and R/BhGLM packages, using a previously studied simulated public epistatic data set.

The general goal of quantitative trait locus (QTL) mapping and association studies is to find certain genomic regions, or QTL, that have nonnegligible contributions to the distribution of quantitative traits (e.g., Xu 2003). In a genome-wide set of markers, the majority of the markers may not link to the QTL and should have zero effects in theory. The ordinary least-squares estimates of marker effects may show large variances. When epistasis is investigated, all possible pairs of marker-by-marker interactions are typically included into the model, which drastically increases the model dimension. This usually leads to an oversaturated model, where the number of marker effects can be many times larger than the sample size (the number of individuals). These problems motivate the implementation of variable selection, including only a subset of important loci in the model and excluding markers irrelevant to the phenotype (e.g., see Broman and Speed 2002; Sillanpää and Corander 2002), and parameter regularization, assigning a penalty term to shrink the marker effects toward zero (Tibshirani 1996). Correspondingly, two classes of Bayesian models (e.g., see O’Hara and Sillanpää 2009) have been developed and used for QTL mapping:

1. Bayesian variable selection models: The first class is based on a two-component mixture prior distribution assigned to each marker/locus effect. In the first component, the distribution of the marker effects has a narrow Gaussian distribution (or point mass) at zero, while in the other component the marker effects can vary in a large range of values. An auxiliary indicator variable is then used to specify which component a marker belongs to. Examples include stochastic search variable selection (SSVS) (George and McCulloch 1993; Yi et al. 2003), Bayes B type methods (Meuwissen et al. 2001; Sillanpää and Bhattacherjee 2005; Habier et al. 2007), and a Bayesian classification approach introduced by Zhang et al. (2005).

2. Bayesian shrinkage models: The second class is based on hierarchical shrinkage modeling, where hierarchical
sparsity-inducing priors are used to shrink unimportant marker effects toward zero. Proposed models include Bayesian adaptive shrinkage (Xu 2003), the Student’s t model (Yi and Xu 2008; Yi and Banerjee 2009), Bayesian LASSO (Park and Casella 2008; Yi and Xu 2008), and extended Bayesian LASSO (Mutshinda and Sillanpää 2010). In class 1, QTL occupancy probabilities and corresponding Bayes factors can be used to judge QTL (Knürr et al. 2011). In class 2, after performing the shrinkage estimation of the marker effects, criteria such as posterior credible intervals (Li et al. 2011) or permutation tests (Xu 2003) can be used to decide the location of the real QTL signals. Note, however, that rigorous decision making between QTL and non-QTL is still an open research problem in Bayesian shrinkage models (Heaton and Scott 2010).

After choosing a particular Bayesian model for QTL analysis, the next issue is to estimate the posterior distribution. Overall, there are two families of methods for reaching this goal. The first family consists of stochastic methods, and the other consists of deterministic point-estimation algorithms. Among the stochastic methods, the most commonly used technique is Markov chain Monte Carlo (MCMC) simulation (Robert and Casella 2004). By using this, we are able to generate a sequence of dependent samples for each unknown parameter from our target (posterior) distribution, and these samples can be used to estimate characteristics, e.g., the posterior mean and the parameter uncertainty around it. So far, most of the proposed Bayesian regression methods in QTL mapping have been based on MCMC methods. However, MCMC, as well as many other sampling-based methods, has huge computational demands when applied to large-scale data problems. Therefore, deterministic methods are usually preferable for large-scale data sets, since they can be faster in terms of computation time. A typical example of such a method is the computation of the posterior mode, or so-called maximum a posteriori (MAP) estimation. Numerical optimization algorithms, including the Newton–Raphson method, coordinate descent, and expectation maximization (EM) are often used for this task (Gelman et al. 2004; Zhang and Xu 2005; Yi and Banerjee 2009; Xu 2010). A third choice, called the variational Bayes (VB) method (Jaakkola and Jordan 2000; Grimmer 2011), can produce a factorized approximated distribution to the targeted posterior in a deterministic manner. It has high computational speed and moreover can give an uncertainty measure for the unknown coefficients.

In this article, we propose three variational Bayes algorithms for Bayesian shrinkage-based QTL and epistasis analysis. The hierarchical shrinkage models that we focus on here are Bayesian adaptive shrinkage (Xu 2003), Bayesian LASSO (Yi and Xu 2008), and extended Bayesian LASSO (Mutshinda and Sillanpää 2010). All of them were originally implemented by MCMC methods. The former two models have been gaining increased interest for QTL analysis, while the latter one was recently developed and therefore needs to be investigated more carefully. To our knowledge, the application of the VB method for quantitative trait locus analysis is very limited. A recent publication is Logsdon et al. (2010), which implements a VB algorithm on the basis of the Bayesian classification model of Zhang et al. (2005) for the genome-wide association study of human data. Another related work is Armagan (2009), in which a variational bridge regression method was proposed, which is closely related to Bayesian LASSO, but produces sparse models using a different mechanism.

The structure of this article is as follows. Methods introduces the Bayesian and frequentist shrinkage-based multiple regression models commonly used in QTL and epistatic analysis. Then we cover the general VB theory and next present the VB algorithm for these shrinkage models. In Example analyses, we demonstrate the efficiency of our VB methods for estimating QTL and epistatic effects by using the North American Barley data (Tinker et al. 1996) and simulated epistasis data from Xu (2007). Comparisons are done from two perspectives: (1) the difference among the three Bayesian shrinkage models and (2) the difference between VB and MCMC computation. Finally, we discuss the strengths and weaknesses of our VB approaches.

Methods

Bayesian shrinkage-based regression models

We consider a standard multiple linear regression model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j + e_i, \quad (1)$$

For QTL mapping, $y_i$ represents the phenotypic value of the $i$th individual in the mapping population, and $x_{ij}$ is the genotypic value of individual $i$ at marker $j$, coded 1 for genotype AA and −1 for BB or AB in a doubled haploid (DH) or a backcross (BC) population resulting from inbred line cross experiments with only two possible genotypes at each marker. $\beta_0$ is the intercept, $\beta_j$ is the effect of locus $j$, and $e_i$ is the random error for individual $i$, following a normal distribution $N(0, \sigma^2_e)$ with mean 0 and unknown variance $\sigma^2_e$. In addition to the main QTL effects, the interaction between some pairs of markers may also contribute to the phenotypic variation. To study this, model (1) can be extended to

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j + \sum_{u<v} \sum_{x_{iu}x_{iv}} \rho_{uv} e_i + e_i, \quad (2)$$

where $\rho_{uv}$ represents the pairwise interaction effect of the marker pair $(u,v)$. Recoding the indexes of Equation 2 properly leads to the simpler expression in (1), where the design matrix $X$ now contains markers and also their interactions.
For example, including ridge regression (Hoerl and Kennard 1970) and in the Introduction, we pointed out that a class of Bayesian shrinkage regression (BSR) approaches has been discussed and used widely for estimating the marker effects. Many of them are related to the frequentist shrinkage approaches, including ridge regression (Hoerl and Kennard 1970) and LASSO (Tibshirani 1996). Ridge regression can be specified as

$$\hat{\beta}_r = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \quad (3)$$

where $\lambda \geq 0$. The $\ell_2$ norm penalty function $\lambda \sum_{j=1}^{p} \beta_j^2$ plays a role of shrinking the regression coefficients toward zero. Shrinkage factor $\lambda$ is a tuning parameter, which determines the level of the shrinkage. $\lambda$ is usually selected externally by a model selection criterion such as cross-validation, Akaike information criterion (AIC), or Bayesian information criterion (BIC). LASSO, on the other hand, uses an $\ell_1$ norm penalty and is defined as

$$\hat{\beta}_l = \arg \min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} | \beta_j | \right\}. \quad (4)$$

In contrast to ridge regression, LASSO is able to shrink some of the marker effects exactly to zero because of the nature of the $\ell_1$ penalty.

Now let us go back to BSR. In BSR, the marker effects are assigned independent hierarchical priors $p(\beta_j | \sigma^2_j) \sim N(\beta_j | 0, \sigma^2_j)$ for each marker. Note that through the article, we use the notation $p(\cdot | \cdot)$ to represent a density function. For example, $N(\cdot | 0, \sigma^2)$ represents the normal density function with mean 0 and variance $\sigma^2$. Further sparsity-inducing prior distributions are chosen for nuisance parameters $\sigma^2_j$ to guarantee that each marker effect has a high probability to be zero on the basis of the assumption of the model. Note that we assume the same amount of shrinkage for all effects in our study regardless of them being main or interaction effects (cf. Sillanpää 2009). In addition, noninformative priors can be used for other unknown parameters in (1), for example, $p(\beta_0) \sim 1$ and $p(\sigma^2) \sim 1/\sigma^2$. These hierarchical prior settings actually play a similar role to the penalty terms in ridge regression or LASSO, which shrink the marker effects toward zero. The choices of the priors for the variances $\sigma^2_j$ usually determine the behavior of the shrinkage. In the following, we list three possible choices, which have been recently proposed and used for mapping quantitative traits:

1. Bayesian adaptive shrinkage (BAS): An improper prior $p(\sigma^2_j) \sim 1/\sigma^2_j$ was used by Xu (2003) for mapping quantitative traits. This can be regarded as an extreme case of an inverse gamma prior $p(\sigma^2_j) \sim \text{Inv-Gamma}(\sigma^2_j | \alpha, \beta)$ with hyperparameters $\alpha$ and $\beta$ (in this article, the parameterization of the inverse gamma, gamma, and exponential distributions follows the book by Gelman et al. (2004)) that together determine the amount of shrinkage and the degree of model sparsity. By setting $\alpha = 0$ and $\beta = 0$, the inverse gamma distribution leads to the improper prior $p(\sigma^2_j) \sim 1/\sigma^2_j$ (Tipping 2001; Figueiredo 2003).

2. Bayesian LASSO (BL): The prior for $\sigma^2_j$ is an exponential distribution $\text{Exp}(\lambda^2/2)$ (with $\lambda > 0$). It has been demonstrated that under this hierarchical structure of the priors, the marginal prior of $\beta_j$ is a Laplace distribution (also called double-exponential distribution) with the density

$$p(\beta_j) = \frac{\lambda}{2} e^{-\lambda | \beta_j |} = \int_{0}^{\infty} N(\beta_j | 0, \sigma^2_j) \text{Exp} \left( \sigma^2_j \frac{\lambda^2}{2} \right) d\sigma^2.$$  

This has been recognized as a Bayesian alternative to the LASSO (Park and Casella 2008). Here $\lambda$ is also used for determining the degree of the model sparsity. Different from LASSO in which $\lambda$ should be selected externally, in the Bayesian LASSO, Gamma($\gamma, \upsilon$) with the predetermined hyperparameters $\gamma$ and $\upsilon$ can be used as a conjugate prior for $\lambda^2$ (other nonconjugate priors are possible), so that $\lambda$ can be updated through the Bayesian updating scheme. The applications of Bayesian LASSO to QTL mapping include Yi and Xu (2008), Xu (2010), and Li et al. (2011).

3. Extended Bayesian LASSO (EBL): Mutshinda and Sillanpää (2010) generalized the regular setting of the Bayesian LASSO prior as $p(\sigma^2_j) \sim \text{Exp}(\sigma^2_j | (\delta^2 \eta^2_j/2))$, where $\delta$ is a global shrinkage factor like $\lambda$ in Bayesian LASSO, and $\eta_j$ is specified individually for each marker. [In our work, the parameterization of $\delta$ and $\eta_j$ differs slightly from that in Mutshinda and Sillanpää 2010, so that here we use $p(\sigma^2_j) \sim \text{Exp}(\sigma^2_j | (\delta^2 \eta^2_j/2))$, which is in agreement with the common parameterization of Bayesian LASSO. More details can be found in the Discussion.] By setting our own shrinkage factors for each individual marker effect, the model has more flexibility for controlling the amount of the local shrinkage and choosing model sparsity. A similar approach called Bayesian adaptive LASSO has been proposed in Sun et al. (2010) with different parameterization. An earlier related frequentist version is adaptive LASSO (Zou 2006).

In the following, we turn to the general theory of variational Bayes approximation.

**A brief introduction to variational Bayes estimation**

The development of VB estimation was motivated by the fact that in the full Bayesian analysis, many posterior
distributions we are interested in are intractable, meaning that it is not possible to derive the marginal likelihood for each parameter analytically. By using VB estimation, we are aiming to find a tractable distribution that can approximate the target posterior. Here, we specially refer to a variational approximation method that seeks a factorized distribution to approximate the original posterior distribution. This VB method was originally adapted from the mean field theory (Parisi 1988) in theoretical physics and has been widely used for various Bayesian inferential problems in the fields of machine learning and signal processing (Beal 2003; Bishop 2006; Šmídl and Quinn 2006). In this section, we briefly go through the foundation of the VB method. More detailed introduction to the topic can be found, for example, in the tutorial of Tzikas et al. (2008). Specifically, let us assume that our target posterior distribution has the form \( p(\theta | \text{data}) \) with unknown parameters \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \). The goal of variational Bayes estimation is to find a tractable distribution \( q(\theta | \text{data}) \) that is “close” to the target posterior. In VB, the approximate distribution \( q \) is supposed to be in a factorized form as

\[
q(\theta | \text{data}) = \prod_{k=1}^{K} q(\theta_k | \text{data}),
\]

where \( \theta_1, \theta_2, \ldots, \theta_K \) is a certain partition of the parameter vector \( \theta \) so that \( K \leq N \) (and each \( \theta_k \) is a block of parameters). Compared to real posterior \( p \), the approximate distribution \( q \) assumes conditional independence of \( \theta_k \) (\( k = 1, \ldots, K \)), so that the posterior correlation among the parameter blocks is ignored. From the perspective of information theory, the “distance” between two probability distributions can be measured by Kullback–Leibler divergence (Kullback and Leibler 1951),

\[
KL(q || p) = \int_{\Theta} q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta,
\]

where \( \Theta \) represents the parameter space of \( \theta \). Note that \( KL(q || p) \geq 0 \), and it equals 0 if and only if the two distributions are the same. In VB, we seek the approximate distribution \( \hat{q}(\theta | \text{data}) \) that satisfies

\[
\hat{q}(\theta | \text{data}) = \arg \min_{q(\theta | \text{data})} KL(q(\theta | \text{data}) || p(\theta | \text{data})),
\]

which means the closest possible approximate distribution to the target posterior \( p(\theta | \text{data}) \). It has been demonstrated (see Šmídl and Quinn 2006) that the minimum of (8) is reached at

\[
\hat{q}(\theta_k | \text{data}) = \frac{\exp \left\{ E_{q(\theta_k | \text{data})} [\ln p(\theta, \text{data})] \right\}}{\int_{\Theta_k} \exp \left\{ E_{q(\theta_k | \text{data})} [\ln p(\theta, \text{data})] \right\} d\theta_k}
\]

\[
\times \exp \left\{ E_{\hat{q}(\theta_k | \text{data})} [\ln p(\theta, \text{data})] \right\} \text{ for } k = 1, 2, \ldots, K,
\]

where \( E_{q(\theta_k | \text{data})} [\ln p(\theta, \text{data})] \) is the expectation of the log-joint distribution with respect to \( q(\theta_k | \text{data}) = \prod_{k \neq j} q(\theta_j | \text{data}) \), the product of distributions of all other partitions of \( \theta \) except \( \theta_k \). It is important to (i) choose suitable (preferably conjugate) prior distributions and (ii) partition the parameter vector \( \theta \) properly to guarantee that each approximate marginal distribution in (10) is tractable. In a convenient case, \( \hat{q}(\theta_k | \text{data}) (k = 1, \ldots, K) \) can be recognized as a standard parametric distribution from (10). Then it is possible to iteratively estimate \( \hat{q}(\theta_k | \text{data}) \) by using the following algorithm:

Step 1: Initialize the approximate marginal distribution for each \( \theta_k \).
Step 2: Update the parameters of each approximate marginal distribution \( \hat{q}(\theta_k | \text{data}) \) in the cyclic manner 1, 2, \ldots, \( K \) until convergence.

Otherwise, if \( \hat{q}(\theta_k | \text{data}) \) cannot be recognized as a standard distribution, numerical integration methods are needed to estimate the denominator part in (9) to formulate \( \hat{q}(\theta_k | \text{data}) \), which is computationally more demanding. In both cases, we are able to obtain approximate marginal distributions for each block of parameters. Thus, we also obtain uncertainty measures for them from the VB approach. (In some cases, the MAP approach can also provide uncertainty measures for the parameters. See Zhang and Xu 2005 as an example. The posterior variance from MAP may be biased downwardly because it is often conditional on other parameters.) For example, if the approximate marginal distribution \( \hat{q}(\theta_k | \text{data}) \) is a standard distribution, then we can easily determine the credible interval for the parameter block \( \theta_k \). In more complicated cases, this requires numerical techniques.

Finally, it is useful to point out that the VB approximation can be understood from the perspective of the following lower bound of the marginal distribution of the data (marginal likelihood); see Šmídl and Quinn (2006). Note that the following relation is satisfied,

\[
\ln p(\text{data}) = \int_{\Theta} q(\theta | \text{data}) \ln \frac{p(\theta, \text{data})}{q(\theta | \text{data})} d\theta
\]

\[
- \int_{\Theta} q(\theta | \text{data}) \ln \frac{p(\theta | \text{data})}{q(\theta | \text{data})} d\theta
\]

\[
= \int_{\Theta} q(\theta | \text{data}) \ln \frac{p(\theta, \text{data})}{q(\theta | \text{data})} d\theta + KL(q(\theta | \text{data}) || p(\theta | \text{data}))
\]

\[
\geq \int_{\Theta} q(\theta | \text{data}) \ln \frac{p(\theta, \text{data})}{q(\theta | \text{data})} d\theta
\]

\[
= L(q(\theta | \text{data})),
\]

where \( L(q(\theta | \text{data})) \) can be interpreted as a lower bound of the log marginal-likelihood function \( \ln p(\text{data}) \), which is supposed to be a constant value, and minimization of \( KL(q(\theta | \text{data}) || p(\theta | \text{data})) \) is equivalent to maximization of \( L(q(\theta | \text{data})) \). The lower bound can be further presented as
\[ L(q(\mathbf{\theta}|\text{data})) = \int_0^1 \prod_{k=1}^K q(\mathbf{\theta}_k|\text{data}) \ln p(\mathbf{\theta}, \text{data}) d\mathbf{\theta} \]
\[ \quad - \int_0^1 \prod_{k=1}^K q(\mathbf{\theta}_k|\text{data}) \sum_{k=1}^K \ln q(\mathbf{\theta}_k|\text{data}) d\mathbf{\theta} \]
\[ = E\prod_{k=1}^K q(\mathbf{\theta}_k|\text{data})[\ln p(\mathbf{\theta}, \text{data})] \]
\[ \quad - \sum_{k=1}^K E_q(\mathbf{\theta}_k|\text{data})[q(\mathbf{\theta}_k|\text{data})]. \]

(12)

It is not difficult to prove that maximizing (12) with respect to \( q(\mathbf{\theta}_k|\text{data}) \) leads to the same equation (Equation 10) that we obtained by minimizing the KL divergence. From another perspective, in the above-mentioned iteration algorithm, after updating each \( q(\mathbf{\theta}_k|\text{data}) \) once, we can then calculate the value of the corresponding lower bound by using (12) and use it as a stopping criterion for the algorithm. Furthermore, the lower bound can also be used as a criterion for model comparison, since it is an approximation to the log marginal-likelihood function of a particular model, which is essentially needed for constructing a Bayes factor (Beal 2003).

In the next sections, we show that the conjugate priors we choose in the hierarchical shrinkage models guarantee that (10) can be easily evaluated.

**VB algorithm for Bayesian shrinkage models**

We are able to formulate posterior approximation algorithms for BAS, BL, and EBL on the basis of the variational Bayes theory described above. In the following, we provide a skeleton of the VB algorithm for the EBL model, which is the most complicated model of the three. The VB procedure can be derived for BAS and BL models in a similar way, and the details are presented in supporting information, File S1.

The likelihood for model (1) can be specified as

\[ p(y|\mathbf{\beta}, \tau_0^2) \propto \prod_{i=1}^n N(y_i|\beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \tau_0^{-2}), \]

(13)

where \( \tau_0^2 = 1/\sigma_0^2 \). In addition, the priors are specified as

\[ p(\beta_0) \propto 1, \]

(14)

\[ p(\tau_0^2) \propto \frac{1}{\tau_0^2}, \]

(15)

\[ p(\beta_j|\tau_j^2) \propto N(\beta_j|0, \frac{1}{\tau_j^2}), \]

(16)

where \( \tau_j^2 = 1/\sigma_j^2 \),

\[ p(\tau_j^2, \eta_j^2) \propto \text{Inv-Gamma}(\tau_j^2|1, \frac{\delta_j^2 \eta_j^2}{2}), \]

(17)

\[ p(\delta_j^2|\gamma, \nu) \propto \text{Gamma}(\delta_j^2|\gamma, \nu), \]

(18)

and

\[ p(\eta_j^2|\psi, \vartheta) \propto \text{Gamma}(\eta_j^2|\psi, \vartheta). \]

(19)

We take each single parameter \( \beta_0, \beta_1, \ldots, \beta_p, \tau_0, \tau_1, \ldots, \tau_p, \delta^2, \eta_1^2, \eta_2^2 \) as a subcomponent of \( \mathbf{\theta} \). Now we seek an approximate distribution to the posterior denoted by

\[ \hat{q}(\theta|y) = \hat{q}(\beta_0|y)\hat{q}(\beta_1|y) \ldots \hat{q}(\beta_p|y)\hat{q}(\tau_0^2|y)\hat{q}(\tau_1^2|y) \ldots \hat{q}(\tau_p^2|y)\hat{q}(\eta_1^2|y) \ldots \hat{q}(\eta_2^2|y). \]

(20)

From Equation 10, we can derive the approximate density of each component as follows:

\[ \hat{q}(\beta_0|y) \text{ is a normal density}, \]

(21)

\[ \hat{q}(\tau_0^2|y) \text{ is a gamma density}, \]

(22)

\[ \hat{q}(\beta_j|y) \text{ is a normal density}, \]

(23)

\[ \hat{q}(\tau_j^2|y) \text{ is an inverse Gaussian density}, \]

(24)

\[ \hat{q}(\delta^2|y) \text{ is a gamma density}, \]

(25)

\[ \hat{q}(\eta_j^2|y) \text{ is a gamma density}. \]

(26)

A detailed description of (21)–(26) is provided in the Appendix. After certain initial values are assigned to the parameters of each distribution \( \hat{q}(\theta|y) \), an iterative algorithm is used to update them successively until convergence. In practice, the lower bound can be calculated in each iteration and be used as a criterion for stopping. Typically, in the \( n \)th iteration, we can calculate \( |L((n)-L^{(n-1)})|/L^{(n)} \). If it is smaller than some predefined threshold (small positive value), then the algorithm should stop. After convergence, we obtain the approximate marginal distributions for each parameter defined in the model.

**Example Analyses**

We use data sets including the well-known North American Barley data (Tinker et al. 1996) and a simulated epistasis data set from Xu (2007) to illustrate the performance of our methods. In the first example, we simulated 50 replicated data sets of phenotypes on the basis of the barley genotype data and empirically assessed the average performance of the methods by presenting results from the replicated data analysis. In the second example, we performed the analysis of epistatic effects on the basis of the simulated data of Xu (2007). In the third example, the sensitivity of the methods was tested with different settings of hyperparameters, using
the barley data with real phenotypes. Differences between VB and MCMC approaches were also investigated. Both barley data and simulated data are available as the supplementary materials of Xu (2007), which can be downloaded from the website http://www.biometrics.tibs.org/.

**Analysis of simulated data replicates**

We used the barley marker data (described below) as a basis for the simulation study. Before the analysis, each missing marker genotype was completed (once) by random draws from Bernoulli($p_{\text{mis}}$), where $p_{\text{mis}}$ is the conditional expectation estimated from flanking markers with known genotypes (see Haley and Knott 1992). We placed QTL positions exactly on the six markers (nos. 2, 20, 40, 60, 80, and 102) with additive effects of $-0.5$, $0.5$, $-0.3$, $0.3$, $-0.8$, and $0.8$, respectively. On this basis, we simulated 50 replicated sets of phenotype data independently, setting the population intercept to 0 and residual variance of normal distribution to 1. By doing so, the average heritability over replications was 0.6881. For each replication, we fitted the BAS, BL, and EBL to the data using both variational Bayes and MCMC estimation algorithms (in the following, we abbreviate the methods as VB-BAS, VB-BL, VB-EBL, MCMC-BAS, MCMC-BL, and MCMC-EBL respectively). In VB-BL, we chose hyperpriors $\lambda^2 \sim \text{Gamma}(1, 0.0001)$, and for VB-EBL, we had $\delta^2 \sim \text{Gamma}(1, 0.0001)$ and $\eta^2 \sim \text{Gamma}(1, 0.0001)$. Note that Gamma(1, 0.0001) is a noninformative prior, which is close to an improper flat uniform prior $\text{Uni}(0, \infty)$, meaning that we have a weak preassumption on the values of $\delta^2$ and $\eta^2$. For comparison, in MCMC-EBL and MCMC-BL, we used the same priors for shrinkage factors as in the VB approaches. In VB, we chose a threshold of $\epsilon = 10^{-6}$, so that if $|L^{(n)} - L^{(n-1)}|/L^{(n)} < \epsilon$ is satisfied, we stopped running the algorithm. Since we obtained an approximate posterior distribution for each parameter from VB, we used the expectation of each posterior distribution as our point estimate. In MCMC, we generated 15,000 samples, where the first 5000 samples were discarded as burn-in, and from the remaining 10,000 samples we stored only every 10th sample to reduce the serial correlation. Next, the posterior mean was calculated and considered as the estimate of the marker effects. Although both VB and MCMC approaches do not shrink any marker effects exactly to zero, they can still be used to guide the variable selection. We used the criterion that if the 95% credible interval contained zero for the estimated coefficient, then the corresponding marker had zero effect (i.e., is not a QTL). The same criterion has been used in Li et al. (2011). For 50 replications, we reported the average number of correctly selected nonzero effects (denoted as C) and that of incorrectly selected nonzero effects (denoted as I) in Table 1. Sometimes, the estimated QTL signals (effects) were detected from the neighboring markers instead of the actual marker positions that were used to generate the phenotypes. Therefore, we also reported the number of correctly selected QTL or their closest neighboring markers (denoted as C extension). In addition, the average mean squared error (MSE) and fivefold cross-validation error (CVE) (see Hastie et al. 2009) were also calculated for each method. The MSE here refers to the average sum of squared errors, defined as $\text{MSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n$, where $\hat{y}_i (i = 1, \ldots, n)$ are the fitted values. We calculated $\hat{y}_i$ by using the formula $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} x_{ij} \hat{\beta}_j$, where $\hat{\beta}_j (j = 0, 1, \ldots, p)$ represented the posterior mean estimates obtained from VB and MCMC methods or the $\ell_1$-penalized least-squares estimates from the LASSO method. Note that for comparison, we also performed a standard LASSO analysis (Tibshirani 1996) by using the coordinate descent algorithm (Friedman et al. 2010) on the same data, where the LASSO shrinkage factor $\lambda$ was chosen so that the minimum of CVE was reached. These results are also presented in Table 1. Finally, the estimated QTL effects averaged over 50 replications for all the methods are shown in Figure 1.

From the perspective of producing the correct number of nonzero effects (i.e., QTL), VB-EBL performs best, because it is able to capture on average 5.50 trait loci, which is close to 6, and at the same time provides accurate estimates of QTL effects. VB-BL also estimates almost 6 trait loci correctly, but it overshinks their QTL effects. VB-BAS tends to identify correct QTL and provides accurate effect estimates for them, but it also gives many spurious signals. Compared to the VB approaches, MCMC-BAS and MCMC-EBL behave equally well in the sense that they accurately estimate the QTL effects. However, the number of correctly identified QTL is much smaller, because the 95% credible intervals estimated from MCMC samples are wider than those from VB. MCMC-
Figure 1  (A–G) In the simulation study, averaged estimated marker effects over 50 replications are plotted (blue lines) along the genome for seven methods. True simulated QTL effects are shown by red asterisks.
Table 2 Average performance of the permutation test to judge QTL signals for VB-BAS, VB-BL, and VB-EBL with 50 replications

<table>
<thead>
<tr>
<th>Method</th>
<th>C</th>
<th>C extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB-BAS</td>
<td>0.4600</td>
<td>0.4600</td>
</tr>
<tr>
<td>VB-BL</td>
<td>6.0000</td>
<td>107.1000</td>
</tr>
<tr>
<td>VB-EBL</td>
<td>3.0800</td>
<td>0.7000</td>
</tr>
</tbody>
</table>

Presented quantities the number of correctly selected QTL (C), the number of incorrectly selected QTL (l), and the number of correctly selected QTL or their closest neighboring markers (C extension).

BL, similarly to VB-BL, tends to overshrink the QTL effects. Compared to the Bayesian methods, standard LASSO is able to shrink marker effects exactly to zero, but it is not able to provide confidence intervals. Thus, we implemented the phenotype permutation method to decide a 95% significance threshold, on the basis of which QTL were selected from the LASSO solution (Churchill and Doerge 1994). To perform a permutation test, the phenotypes were reshuffled to destroy the association between the phenotypes and the markers for 1000 runs. For each run, we estimated the marker effects and used the largest absolute effect to construct an empirical distribution. We used its 95% quantile as a threshold, so that a marker with the absolute effect larger than the threshold was considered to be a QTL. Results showed that 2.32 QTL on average were detected. Furthermore, we also performed the permutation test for the three VB methods, and the results are summarized in Table 2. It seems that the permutation test performed well with VB-EBL. It correctly detected 3.72 QTL on average, which is less than the number of QTL detected by the credible intervals, but it also provided less falsely selected QTL. For VB-BAS, the estimated threshold was too high so that only 0.46 markers were detected to be the QTL on average. For VB-BL, the threshold was too low, and we obtained too many false positive signals. Moreover, MSE is often used to measure the goodness-of-fit of the model to the data. From the results, it can be seen that MCMC-BL and VB-BAS have smaller MSE on average than others, meaning that they fit the data better. On the other hand, cross-validation provides a measure of predictive ability for a method. Standard LASSO, MCMC-EBL, VB-BL, and MCMC-BAS approaches tend to give smaller CVE, meaning that they have good predictive ability. MCMC-BL and VB-BAS produced the highest CVEs, indicating that these two methods may suffer from overfitting in this replication study.

Analysis of simulated data to estimate epistatic effects

A backcross data set with 600 individuals was simulated in Xu (2007). In this data set, 121 markers were evenly distributed through a single chromosome of 1800 cM. Simulated QTL with main effects were exactly placed on 9 markers, and 13 marker pairs were simulated to have interaction effects. Phenotypes were simulated as the cumulative sum of all main and pairwise interaction effects. The population mean and residual variance of normal distribution used in the simulation were 5.0 and 10.0, respectively (the resulting heritability was 0.9065). See Xu (2007) for more details on these data.

We tested the efficiency of VB methods to estimate epistatic effects by using simulated backcross data from Xu (2007). Note that the same linear regression model presented in (1) can also be used here. Design matrix $X$ should include not only 121 markers, but also $(121^2 - 121)/2 = 7260$ pairwise interaction terms between each possible pair of markers, which leads to an ill-posed ($P \gg n$) regression problem. We implemented VB-BAS, VB-BL, and VB-EBL as well as MCMC-BAS, MCMC-BL, and MCMC-EBL (with the same gamma priors for $\delta$ and $\eta^2$ as in the previous example analysis) on these data. Here we use the same priors for the main and interaction effects. The 95% credible intervals-based criterion is also considered as the tool to judge the QTL for each method.

Results are summarized in Table 3 and Figures 2 and 3. For methods including VB-BAS, MCMC-BAS, VB-EBL, and MCMC-EBL, we conclude that their performance is good in general, because they are able to detect almost all the correct loci with main effects as well as the interaction terms, and the estimated QTL effects are also accurate. VB-EBL, MCMC-EBL, and MCMC-BAS fail to find an interaction effect at marker pair (41, 61). In addition, VB-BL does not find an effect at pair (21, 22). MCMC-EBL produced a quite accurate estimate of 0.7015 at (21, 22) that is close to the true effect size 1, but the pair (21, 22) is not recognized as a QTL by the interval test. Compared to the former three methods, VB-BAS is able to find all the important signals, but tends to give many spurious signals in addition to the correct ones. Finally, the results obtained from VB-BL and MCMC-BL are not satisfactory. They can detect only a few major QTL, and the estimated QTL effects in these positions are not accurate. Similar analysis has been done by using empirical Bayes, LASSO, penalized maximum likelihood (Zhang and Xu 2005), and SSVS in Xu (2007). Results obtained from our variational Bayes approaches including VB-EBL and VB-BAS are also competitive with the results published earlier.

Finally, we employed still two other efficient methods for detecting epistatic effects and compared them to our methods. The first one is an EM algorithm for estimating epistatic effects introduced by Yi and Banerjee (2009). In summary, a Bayesian hierarchical model was proposed with priors $\beta_j|\sigma^2 \sim N(0, \sigma^2)$ and $\sigma^2 \sim \text{Inv-}\chi^2(v_j, s^2_j)$ ($j = 1, \ldots, p$). Note that an inverse-$\chi^2$ distribution Inv-$\chi^2(v, s^2)$ is equivalent to an inverse gamma distribution Inv-$\Gamma(v/2, vs^2/2)$. If both $v$ and $s^2$ were zero, we exactly obtain a BAS model. In our example, we chose $v = 0.01$ and $s^2 = 0.0001$, as Yi and Banerjee (2009) suggested. The method has been implemented in the recently developed R package BhGLM (http://www.ssg.uta.edu/bhglm/), which takes the advantage of R’s built-in function glm. We found that the method cannot be directly applied to this data set due to the memory limitation of the computer. Alternatively, we used a model search strategy (with thresholds $t_1 = 10^{-8}$ and $t_2 = 0.1$) also introduced...
in Yi and Banerjee (2009), which gradually filters out the markers with negligible effects and therefore reduces the dimensionality of the model. The second method is a MCMC-based model-finding algorithm of Yi et al. (2005), which has been implemented in the R package qtlbim (see Yandell et al. 2007). We set the expected number of QTL with main effects (mean.nqtl) to be 9, and the expected total number of QTL with both main and epistatic effects (mean.nqtl) to be 20 on the basis of the number of main and epistatic effects detected by VB-EBL. The results are shown in Figure 4. It seems that R/BghGLM does not provide as good results as VB-EBL. First of all, the locations of two main QTL, which were found at 88 and 113, are biased from the true locations 91 and 101. Second, many spurious QTL signals were detected in R/BghGLM analysis. These may be mainly caused by the use of the model search strategy. A simultaneous estimation of all the markers and the interaction terms may lead to better results. On the other hand, R/qtlbim seems to work nicely. Since the Bayesian model behind R/qtlbim assigned an indicator variable to each marker, an estimate of the Bayes factor can be derived from the MCMC samples and be used to judge the QTL positions. Results in Figure 4C indicate the markers with large values of Bayes factors were located either exactly at the positions of true main QTL or close to them. Furthermore, R/qtlbim was also able to correctly detect most of the marker pairs with true epistatic effects, as illustrated in Figure 4D. Like VB-EBL, R/qtlbim failed to detect the epistatic effects at marker pairs (21, 22) and (41, 61). In addition, it produces a few false positive signals. We have also found that the performance of R/qtlbim relies on the choice of priors including main.nqtl and mean.nqtl. As we have shown, it is possible to set those priors on the basis of results from our VB methods.

**Real barley data analysis**

The mapping population consists of 145 doubled haploid lines (n = 145), each grown in a range of environments. A total of 127 markers were genotyped, covering 1270 cM of the barley genome, with the average distance between markers of 10.5 cM. Seven traits including yield, heading, maturity, height, lodging, kernel weight, and test weight were measured for each line. Here we selected the kernel weight as the quantitative trait used in the analysis. Before the analysis, the phenotype values for each line in the traits were averaged over the environments. Missing marker genotypes were imputed in the same way as in the first example analysis. The VB methods based on the three models as well as the corresponding MCMC approaches were used for the analysis. To test the prior sensitivity of VB-GL and VB-EBL, we implemented VB-GL with different combinations of hyper-parameters as (γ1, ν1) = (0.0001, 0.0001), (γ2, ν2) = (1, 0.0001), and (γ3, ν3) = (1, 1), respectively. Correspondingly, for VB-EBL, we chose (γ1, ν1) = (ψ1, θ1) = (0.0001, 0.0001), (γ2, ν2) = (ψ2, θ2) = (1, 1), respectively. In MCMC-EBL and MCMC-GL, we used the same priors as in the corresponding VB approaches.

### Table 3 Estimated main and interaction effects from VB-BAS, MCMC-BAS, VB-GL, MCMC-GL, VB-EBL, and MCMC-EBL for the backcross data originally simulated by Xu (2007)

<table>
<thead>
<tr>
<th>Markers</th>
<th>True effects</th>
<th>VB-BASa</th>
<th>MCMC-BAS</th>
<th>VB-GLb</th>
<th>MCMC-GLb</th>
<th>VB-EBL</th>
<th>MCMC-EBLc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1)</td>
<td>4.47</td>
<td>4.5419</td>
<td>4.4633</td>
<td>2.7849</td>
<td>3.6481</td>
<td>4.4240</td>
<td>4.4529</td>
</tr>
<tr>
<td>(21, 21)</td>
<td>3.16</td>
<td>3.1603</td>
<td>3.1408</td>
<td>0.8826</td>
<td>1.6565</td>
<td>3.1502</td>
<td>3.1287</td>
</tr>
<tr>
<td>(31, 31)</td>
<td>2.24</td>
<td>2.3097</td>
<td>2.2777</td>
<td>0.6063</td>
<td>1.2301</td>
<td>2.2495</td>
<td>2.2548</td>
</tr>
<tr>
<td>(51, 51)</td>
<td>1.58</td>
<td>1.3123</td>
<td>1.3294</td>
<td>0.0530</td>
<td>0.2445</td>
<td>1.3090</td>
<td>1.2606</td>
</tr>
<tr>
<td>(71, 71)</td>
<td>1.58</td>
<td>1.7562</td>
<td>1.5349</td>
<td>0.0614</td>
<td>0.3275</td>
<td>1.5042</td>
<td>1.5099</td>
</tr>
<tr>
<td>(91, 91)</td>
<td>1.10</td>
<td>1.0180</td>
<td>0.9041</td>
<td>0.0233</td>
<td>0.1589</td>
<td>0.8849</td>
<td>0.9010</td>
</tr>
<tr>
<td>(101, 101)</td>
<td>1.10</td>
<td>1.2624</td>
<td>1.1763</td>
<td>0.0727</td>
<td>0.3102</td>
<td>1.1819</td>
<td>1.1705</td>
</tr>
<tr>
<td>(111, 111)</td>
<td>0.77</td>
<td>0.4752</td>
<td>0.6844</td>
<td>0.0169</td>
<td>0.0611</td>
<td>0.6737</td>
<td>0.5742</td>
</tr>
<tr>
<td>(121, 121)</td>
<td>0.77</td>
<td>0.4799</td>
<td>0.5484</td>
<td>0.0490</td>
<td>0.1638</td>
<td>0.5565</td>
<td>0.4929</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>1.00</td>
<td>0.6421</td>
<td>0.6958</td>
<td>0.0199</td>
<td>0.1000</td>
<td>0.6545</td>
<td>0.6147</td>
</tr>
<tr>
<td>(2, 119)</td>
<td>3.87</td>
<td>3.4860</td>
<td>3.8404</td>
<td>1.0283</td>
<td>1.7288</td>
<td>3.8556</td>
<td>3.8195</td>
</tr>
<tr>
<td>(10, 91)</td>
<td>1.30</td>
<td>1.5121</td>
<td>1.3167</td>
<td>0.0432</td>
<td>0.1641</td>
<td>1.2087</td>
<td>1.2381</td>
</tr>
<tr>
<td>(15, 75)</td>
<td>1.73</td>
<td>1.2712</td>
<td>1.5639</td>
<td>0.0810</td>
<td>0.1761</td>
<td>1.5027</td>
<td>1.4724</td>
</tr>
<tr>
<td>(20, 46)</td>
<td>1.00</td>
<td>0.7548d</td>
<td>0.9338</td>
<td>0.0200</td>
<td>0.0832</td>
<td>0.8834</td>
<td>0.8782</td>
</tr>
<tr>
<td>(21, 22)</td>
<td>1.00</td>
<td>0.7588</td>
<td>0.9808</td>
<td>0.0139</td>
<td>0.0756</td>
<td>0.0100</td>
<td>0.0715</td>
</tr>
<tr>
<td>(26, 91)</td>
<td>1.00</td>
<td>1.2658</td>
<td>1.2785</td>
<td>0.0449</td>
<td>0.1532</td>
<td>1.2263</td>
<td>1.2372</td>
</tr>
<tr>
<td>(41, 61)</td>
<td>0.71</td>
<td>0.3119d</td>
<td>0.0281</td>
<td>0.0038</td>
<td>0.0000</td>
<td>0.0120</td>
<td>0.0502</td>
</tr>
<tr>
<td>(56, 91)</td>
<td>3.16</td>
<td>2.8543d</td>
<td>3.0684d</td>
<td>0.6414d</td>
<td>1.0039d</td>
<td>3.0348d</td>
<td>3.0206d</td>
</tr>
<tr>
<td>(65, 85)</td>
<td>2.24</td>
<td>2.2781</td>
<td>2.4525</td>
<td>0.3207</td>
<td>0.8113</td>
<td>2.4204</td>
<td>2.4279</td>
</tr>
<tr>
<td>(86, 96)</td>
<td>0.89</td>
<td>0.8318</td>
<td>1.0019</td>
<td>0.0211</td>
<td>0.0624</td>
<td>0.9601</td>
<td>0.9472</td>
</tr>
<tr>
<td>(101, 105)</td>
<td>1.00</td>
<td>1.0368</td>
<td>1.0223</td>
<td>0.0516</td>
<td>0.1111</td>
<td>0.9754</td>
<td>0.9136</td>
</tr>
<tr>
<td>(111, 121)</td>
<td>2.24</td>
<td>2.0512</td>
<td>2.3086</td>
<td>0.4615</td>
<td>1.0614</td>
<td>2.3373</td>
<td>2.3289</td>
</tr>
</tbody>
</table>

*In addition to those reported, VB-BAS detects 307 spurious QTL signals. For both VB-GL and MCMC-GL, markers (or marker pairs) (1, 1), (21, 21), (31, 31), (2, 119), (56, 90), (65, 85), and (111, 121) are judged as QTL on the basis of the 95% credible interval test. For MCMC-EBL, marker pair (21, 22) is not judged as a QTL on the basis of the 95% credible interval test. The estimated QTL effects are not shown exactly in the position where the true QTL effects are simulated, but found in their neighboring markers.

The estimated QTL effects are not shown exactly in the position where the true QTL effects are simulated, but found in their neighboring markers.
Together with VB-BAS, results of VB estimation are presented in Figure 5. Additionally, results of MCMC estimation are shown in Figure 6. We see that the VB-EBL is quite sensitive to the choices of the hyperparameters. By our preferable choice of shape parameter as 1 and rate parameter as 0.0001, VB-EBL tends to produce a relatively sparse model compared to the other two choices. On the other hand, VB-BL seems to be less sensitive to the choice of the hyperparameters. We also note the difference between the results obtained from VB and MCMC methods. For example, in the BAS model, the MCMC approach leads to a sparser model than the VB approach. Another interesting feature is that the credible interval estimated from VB methods tends to be narrower than the one estimated from the MCMC samples based on the same shrinkage model. Figure 7 shows an example, where we plot the estimated effects of

**Figure 2** Estimated main QTL effects (blue line) compared to simulated main QTL effects (red asterisks) from (A) VB-BAS, (B) MCMC-BAS, (C) VB-BL, (D) MCMC-BL, (E) VB-EBL, and (F) MCMC-EBL in data from Xu (2007).
markers (in blue) which are announced to be significant (95% credible intervals do not contain zero) from VB-EBL. The corresponding posterior means and 95% C.I.’s estimated by MCMC (in red) are also shown in Figure 7. Clearly, all the MCMC-derived 95% C.I.’s contain VB-derived 95% C.I.’s. Consequently, from VB, we obtain 12 QTL by interval inference, but from MCMC, only 3 of them including markers 12, 43, and 102 are claimed to be significant. This implies that, in practice, the VB inference based on intervals may tend to give more false positive results. Alternatively, we performed a permutation test with VB-EBL, and three markers, 2, 12, and 102 were found to be significant. These markers are located close to the QTL signals reported in Tinker et al. (1996).

Discussion

We propose a variational Bayes approximation-based algorithm for Bayesian shrinkage analysis of multiple-QTL mapping. Because of the usage of the conjugate priors, iterative algorithms can be easily derived for hierarchical Bayesian shrinkage models including BAS, BL, and EBL to
update the approximate marginal distribution of each parameter. Compared to MCMC methods, which are used to generate dependent samples from the target posterior distribution directly, VB seeks an approximate distribution in a factorized form that has minimum KL divergence to the target posterior. Therefore, it is not surprising that for a particular model, the posterior mean estimated by MCMC can be different from that estimated by VB. Another feature is that the credible intervals estimated by VB methods are narrower than those of MCMC. Therefore, if the credible intervals are used to judge QTL, VB tends to give more false positive signals. On the other hand, MCMC-based shrinkage methods tend to provide too large credible intervals arguably due to the collinearity between loci, and a criterion based on these may fail to find some true QTL signals. Therefore, we suggest that a credible intervals-based criterion should be used for QTL detection only in the first stage. A further method may be needed to judge QTL in a more accurate way in the second stage. We found that a permutation-based test might be a suitable tool to control false positive findings under the VB-EBL model, but not for VB-BL and VB-BAS. However, performing such a test often requires huge computation time and is not preferable for the high-dimensional data. As Heaton and Scott (2010) pointed out, how to select the significant variables under a Bayesian shrinkage procedure remains an open problem and needs to be investigated in further studies. Compared to the MCMC approaches, one major advantage of VB approaches is their

Figure 4 (A and B) The absolute value of estimated main and epistatic effects from the hierarchical generalized linear models (in B, the blue lines on the right represent the estimated interacting effects, and the red lines on the left represent the simulated interacting effects). (C) The Bayes factor (BF) (in logarithmic scale) for main effects obtained by R/qtlbim. The red vertical lines indicate the true QTL positions. The horizontal dashed line indicates a QTL threshold $2 \ln BF \approx 2.2$ (corresponding to BF = 3) as suggested in Kass and Raftery (1995). (D) The Bayes factors (in logarithmic scale) of epistatic effects on the upper triangle of a two-dimensional grid. The color scale (from red to blue) corresponds to BF with different values (from high to low). The black dots indicate the locations of true epistatic effects.
Figure 5 Estimated marker effects are plotted along the genome from the North American Barley study for (A, C, and E) VB-EBL with different sets of hyperparameters, (B, D, and F) VB-BL, and (G) VB-BAS.
Figure 6 Estimated marker effects are plotted along the genome from the North American Barley study for (A, C, and E) MCMC-EBL with different sets of hyperparameters, (B, D, and F) MCMC-BL, and (G) MCMC-BAS.
Figure 7 A comparison between 95% credible intervals estimated from VB (in blue) and MCMC (in red).

computational efficiency. Note that the computation time for a single iteration in VB is approximately identical to a Gibbs sampling step. On the basis of our empirical experiments, VB algorithms require ~100–1500 iterations until convergence. On the other hand, when MCMC is used, we usually need to draw >10,000 MCMC samples to obtain good posterior summaries. Therefore, compared to MCMC, VB methods require much less computation time. Table 4 presents a comparison of the computation time of both VB and MCMC in our second and third example analyses. All the methods were implemented in MATLAB on a regular desktop with a 3.00-GHz Intel Pentium 4 processor and 3 GB RAM. The MATLAB source code is available in File S2.

The VB method we consider here can be regarded as an extension of the well-known EM algorithm and can be applied to a broader class of Bayesian models. In an EM algorithm, when the hidden variables interact with each other and their joint full conditional distribution is not analytically available, the variational approximation may be applied as an alternative to proceed to the E-step. This is then called the VB-EM algorithm (Beal 2003). Furthermore, if the parameters, with respect to which the objective function is maximized, are considered as the hidden variables so that only the E-step is remaining, then we exactly obtain the VB algorithm mentioned in this article. More information about the connection between VB and EM algorithms is provided in File S1.

This article discusses three Bayesian shrinkage models, including BL, EBL, and BAS. BL and EBL can be regarded as the same model class, which uses double exponential prior distributions of marker effects. Compared to BL, an improvement in EBL is to assign a marker-specific shrinkage factor $\eta_j$ to each marker in addition to a global shrinkage factor $\delta$. Mutshinda and Sillanpää (2010) pointed out that, by using this marker-specific shrinkage factor, EBL is able to relax the penalty for the important marker effects and meanwhile push the unimportant marker effects strongly toward zero. From the results of our example analyses, it can be seen that VB-EBL outperformed VB-BL. In our simulated examples, VB-EBL gave accurate estimates of the marker effects and meanwhile produced a relatively sparse model. On the other hand, VB-BL tended to shrink the large marker effects too much, so that the estimation was biased. BAS, a third model we focused on, comes from a different model class, where the Student’s $t$ prior is assigned for the marker effects. In the simulation studies, we found that VB-BAS can produce accurate estimates of the marker effects, but the resulting model was not sparse compared to VB-EBL [with our default hyperparameters $(\gamma, \nu) = (1, 0.0001)$], meaning that some spurious signals were present. Therefore, VB-BAS may not be a preferable method for QTL mapping, in which we are more interested in detecting QTL with relatively large effects on the traits.

The Bayesian LASSO method has also been used for predicting phenotypic values and breeding values (De Los Campos et al. 2009; Legarra et al. 2011). In the first replication study, we used cross-validation to measure the predictive ability of our proposed methods, and we found that the performance of VB-EBL is competitive to LASSO, which has been widely used for prediction problems in many areas. However, the predictive performance of a certain method depends on the property of data sets, i.e., the distribution of QTL effects (Daetwyler et al. 2010; Clark et al. 2011). The general predictive abilities of our VB methods need to be investigated in more complicated data sets.

In addition, it is necessary to point out the difference between our EBL and BL models and other approaches. Compared to Mutshinda and Sillanpää (2010), we used different parameterizations and priors of the shrinkage factors in the EBL and BL models. Taking the EBL model as an example, in Mutshinda and Sillanpää (2010), $\delta$ and $\eta_j$ were treated as parameters, and they were assigned the uniform priors $\text{Uni}(0, 100)$. However, in our VB approaches, $\delta^2$ and $\eta_j^2$ are treated as parameters, and conjugate priors $\delta^2 \sim \text{Gamma}(\gamma, \nu)$ and $\eta_j^2 \sim \text{Gamma}(\psi, \theta)$ are used to guarantee the convenience of the computation. For comparison, here we used the same gamma priors for MCMC as in the VB approaches. However, on the basis of our empirical experiments, EBL is sensitive to the hyperparameters with the

<table>
<thead>
<tr>
<th>Data size</th>
<th>VB-BAS</th>
<th>VB-BL</th>
<th>VB-EBL</th>
<th>MCMC-BAS</th>
<th>MCMC-BL</th>
<th>MCMC-EBL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 127, n = 145$</td>
<td>4.05</td>
<td>0.48</td>
<td>0.72</td>
<td>111.50</td>
<td>112.85</td>
<td>114.66</td>
</tr>
<tr>
<td>$p = 7260, n = 600$</td>
<td>304.28</td>
<td>182.00</td>
<td>53.43</td>
<td>6054.63</td>
<td>6278.37</td>
<td>6511.36</td>
</tr>
</tbody>
</table>

Table 4 Comparison of the approximate computation time (scaled in seconds) of VB and MCMC methods for the barley ($p = 7260, n = 145$) and simulated data by Xu (2007) ($p = 7260, n = 600$)
gamma priors in MCMC, and the use of the uniform priors suggested by Mutshinda and Sillanpää (2010) might be a good choice to avoid tuning.

In BL and EBL, another remaining issue is “unimodality”. To guarantee unimodality of the posterior, Park and Casella (2008) suggested the incorporation of the global residual variance $\sigma^2$ into the prior distribution of marker effects as $\beta_j \sim N(0, \sigma^2 \lambda_j^2)$. They demonstrated that, by doing this, the posterior distribution has a unique mode (maximum value), and the corresponding MCMC algorithm may converge faster. However, in VB estimation, instead of the posterior mode, we seek the minimum value of the KL divergence between the factorized approximate distribution and the original posterior. Thereby, the model suggested by Park and Casella (2008) may not be optimal for our variational Bayes approach. In fact, we found that the VB algorithm based on the Park and Casella model even slowed down the convergence. Also note that the uniqueness of the minimization (i.e., convergence to a global minimum) of the KL divergence in VB estimation is not guaranteed in general (see Šmídl and Quinn 2006). Further investigation may be needed to check for uniqueness of the VB approximation.

Finally, we point out some weaknesses of our current VB approach, which can potentially be improved in future work. First, we derived the approximate marginal distribution for each marker effect $\beta_j$ separately. During the approximation, the posterior correlations among marker effects $\beta_j$ are lost. It may be beneficial to take all marker effects as a block $\mathbf{B} = [\beta_1, \ldots, \beta_p]$ and update them simultaneously. This may help to reduce the minimum value of KL divergence and may give a better approximation to the target posterior distribution. In practice, this involves the inversion of a $p \times p$ matrix $E[\tau_2^2]X^TX + \text{diag}(E(\tau^2))^{-1}$, where $E(\tau^2) = [E(\tau_1^2), \ldots, E(\tau_p^2)]$, which is demanding in terms of computation and storage when $p$ is large. This problem can be potentially solved by applying “Woodbury matrix identity” (Woodbury 1950), which can be applied to convert the $p \times p$ matrix into an $n \times n$ one, so that the dimension of the matrix is reduced significantly. Second, in VB-EBL and VB-BL, we used gamma priors for shrinkage factors and chose shape parameters to be sensitive to the choice of the gamma parameters (see our third example analysis). A third solution to this problem might be that we perform a prior sensitivity analysis using cross-validation for new data to give optimal models on the basis of different data sets. Second, we have found that in the formula to update $q(\tau^2_j | y) (j = 1, \ldots, p)$ (in step VI in the Appendix), if we heuristically replace $E[1/\tau_j^2]$ by $1/E[\tau_j^2] = 1/\mu$, VB-EBL can produce equally good results, and these slight modifications seem to make VB less sensitive to hyperparameters, although they do not have any theoretical justification. Logsdon et al. (2010) also made some slight modifications in their VB algorithm to maintain the ease of computation, and they stated that those modifications did not significantly affect the results. However, in this article, we have presented only the results from standard VB.

Acknowledgments

We are grateful to Petri Koistinen for giving constructive comments on the manuscript. This work was supported by the Finnish Graduate School of Population Genetics and by research grants from the Academy of Finland and the University of Helsinki’s Research Funds.

Literature Cited


Communicating editor: F. Zou
Appendix: VB Estimation Algorithm for Extended Bayesian LASSO

First, the logarithm of the joint distribution of parameters and data is

\[ \ln p(\mathbf{y}, \mathbf{\theta}) = p(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \ldots, \tau_p^2; \mathbf{y}) \]

\[ = - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \ln \Gamma(\frac{\eta_0}{2}) + \frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \frac{\tau_j^2}{\sigma^2} \]

\[ + \left[ \frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \frac{\tau_j^2}{\sigma^2} \right] + (\gamma - 1) \ln \sigma^2 - \frac{1}{2} \sigma^2 \]

\[ = \left[ \frac{1}{2} \sum_{i=1}^{n} \ln \tau_i^2 - \frac{1}{2} \sum_{i=1}^{n} \frac{\tau_i^2}{\sigma^2} \right] + C. \]

(A1)

where \( C \) is a constant. Next, we start to derive the VB marginal distributions \( q(\mathbf{\beta}_0 | \mathbf{y}) \) for parameters \( \beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2, \delta^2, \eta_1^2, \ldots, \eta_p^2 \) on the basis of Equation 10.

I. Derivation of \( q(\mathbf{\beta}_0 | \mathbf{y}) \)

By keeping only terms containing \( \beta_0 \), we obtain

\[ \ln q(\mathbf{\beta}_0 | \mathbf{y}) \propto E_q(\mathbf{\beta}_0 | \mathbf{y}) [\ln p(\mathbf{\theta}, \mathbf{y})] \]

\[ = - \frac{1}{2} E[\tau_0^2] \left\{ n\beta_0^2 - 2\beta_0 \sum_{i=1}^{n} y_i - \sum_{j=1}^{p} x_{ij} \left[ E[\beta_j] \right] \right\} + C, \]

(A2)

where \( q(\mathbf{\theta} / \mathbf{\beta}_0 | \mathbf{y}) \) represents the product of all other approximate marginal distributions except \( q(\mathbf{\beta}_0 | \mathbf{y}) \). Here \( E[\cdot] \) is a simplified notation of the expectation of the parameter \( \cdot \) with respect to its approximate marginal distribution \( q(\mathbf{\cdot} | \mathbf{y}) \) [i.e., \( E[\tau_0^2] \) is a simplified notation of \( E_q(\tau_0^2 | \mathbf{y}) = \int q(\tau_0^2 | \mathbf{y}) d\tau_0^2 \), and \( C \) represents those terms that do not contain \( \beta_0 \). We can recognize from (A2) that \( q(\mathbf{\beta}_0 | \mathbf{y}) \) is a normal distribution with mean

\[ E[\beta_0] = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij} E[\beta_j] \right) \]

(A3)

and variance

\[ \text{Var}[\beta_0] = \frac{1}{n E[\tau_0^2]} \]

(A4)

To update \( E[\beta_0] \) and \( \text{Var}[\beta_0] \), we need to calculate \( E[\beta_j] \) and \( E[\tau_j^2] \) on the basis of approximate marginal distributions \( \hat{q}(\beta_j | \mathbf{y})(j = 1, \ldots, p) \) and \( \hat{q}(\tau_j^2 | \mathbf{y}) \), respectively, which are shown in the following.

II. Derivation of \( q(\beta_j | \mathbf{y})(j = 1, \ldots, p) \)

Following a similar procedure to that in (1), we have

\[ \ln q(\beta_j | \mathbf{y}) \propto E_q(\beta_j | \mathbf{y}) [\ln p(\mathbf{\theta}, \mathbf{y})] \]

\[ = - \frac{1}{2} E[\tau_j^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2] \beta_j^2 \]

\[ + \beta_j E[\tau_j^2] \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij} + C, \]

(A5)

showing that \( q(\beta_j | \mathbf{y}) \) is also a normal distribution with mean

\[ E[\beta_j] = \left( E[\tau_j^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2] \right)^{-1} \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij} \]

(A6)

and variance

\[ \text{Var}[\beta_j] = (E[\tau_j^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1}. \]

(A7)

In addition, the expectation of \( \beta_j^2 \) (second moment) can be calculated by \( E[\beta_j^2] = E[\beta_j]^2 + \text{Var}[\beta_j] \), which is needed later.

III. Derivation of \( q(\tau_j^2 | \mathbf{y}) \)

We have

\[ \ln q(\tau_j^2 | \mathbf{y}) \propto E_q(\tau_j^2 | \mathbf{y}) [\ln p(\mathbf{\theta}, \mathbf{y})] \]

\[ = (a_1 - 1) \ln \tau_j^2 - \frac{b_1 \tau_j^2}{2} + C, \]

(A8)

where

\[ a_1 = \frac{n}{2}, \]

(A9)

and

\[ b_1 = \frac{\sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik})^2 + \sum_{i=1}^{n} \sum_{j \neq i} x_{ij}^2 \text{Var}[\beta_j]}{2} \]

(A10)

Here \( q(\tau_j^2 | \mathbf{y}) \) is recognized as a gamma distribution with parameters shape parameter \( a_1 \) and rate parameter \( b_1 \). In addition, we need to calculate the mean \( E[\tau_j^2 | \mathbf{y}] = a_1 / b_1 \) to update other approximate distributions.

IV. Derivation of \( q(\tau_j^2 | \mathbf{y})(j = 1, \ldots, p) \)

\[ \ln q(\tau_j^2 | \mathbf{y}) \propto E_q(\tau_j^2 | \mathbf{y}) [\ln p(\mathbf{\theta}, \mathbf{y})] \]

\[ = - \frac{3}{2} \ln \tau_j^2 - \frac{1}{2} E[\beta_j^2] \tau_j^2 - \frac{1}{2} \frac{E[\tau_j^2] E[\eta_j^2]}{\tau_j^2} + C. \]

(A11)

Here \( q(\tau_j^2 | \mathbf{y}) \) is recognized as an inverse Gaussian distribution (Seshadri 1999) with parameters
We also need the expectation

\[ \mu = \sqrt{\frac{E[\delta^2 | E[\eta_j^2]]}{E[\beta_j^2]}} \]  

(A12)

and

\[ \lambda_j = E[\delta^2 | E[\eta_j^2]] \].  

(A13)

Note that a general form of the inverse Gaussian density is

\[ p(x | \mu, \lambda) = \left( \frac{\lambda}{2\pi x^3} \right)^{1/2} \exp \left\{ -\frac{\lambda(x-\mu)^2}{2\mu^2 x^2} \right\} (x > 0, \mu > 0, \lambda > 0). \]  

(A14)

To update other approximate distributions, we need to obtain the expectation of \( \tau_j^2 \) as \( E[\tau_j^2] = \mu \), as well as the expectation of \( 1/ \tau_j^2 \) as \( E[1/ \tau_j^2] = 1/ \mu + 1/ \lambda_j \).

V. Derivation of \( \hat{q}(\delta^2 | y) \)

\[
\ln \hat{q}(\delta^2 | y) \propto E_{\hat{q}(\theta/\delta^2 | y)}[\ln p(\theta, y)]
\]

\[ = (p + \gamma - 1) \ln \delta^2 - \left( \frac{1}{2} \sum_{j=1}^{p} E[\eta_j^2] E \left[ \frac{1}{\tau_j^2} \right] + \nu \right) \delta^2 + C. \]  

(A15)

Here \( \hat{q}(\delta^2 | y) \) is recognized as a gamma distribution with shape parameter

\[ a_2 = p + \gamma \]  

(A16)

and rate parameter

\[ b_2 = \frac{1}{2} \sum_{j=1}^{p} E[\eta_j^2] E \left[ \frac{1}{\tau_j^2} \right] + \nu. \]  

(A17)

We also need the expectation \( E[\delta^2] = a_2 / b_2 \).

VI. Derivation of \( \hat{q}(\eta_j^2 | y) \) (\( j = 1, \ldots, p \))

\[
\ln \hat{q}(\eta_j^2 | y) \propto E_{\hat{q}(\theta/\eta_j^2 | y)}[\ln p(\theta, y)]
\]

\[ = \psi \ln \eta_j^2 - \left( \frac{1}{2} E[\delta^2] E \left[ \frac{1}{\tau_j^2} \right] + \delta \right) \eta_j^2 + C. \]  

(A18)

Thus, \( \hat{q}(\eta_j^2 | y) \) is also a gamma distribution with shape parameter

\[ a_3 = 1 + \psi \]  

(A19)

and rate parameter

\[ b_3 = \frac{1}{2} E[\delta^2] E \left[ \frac{1}{\tau_j^2} \right] + \delta. \]  

(A20)

The expectation \( E[\eta_j^2] = a_3 / b_3 \) also needs to be calculated.

To update the approximate distributions in Appendix sections I–VI successively, we need to first assign initial values to the required expectations. In practice, we use \( E[\beta_j] = 0 \) (\( j = 0, 1, \ldots, p \)), \( E[\tau_j^2] = 2 \) (\( j = 0, 1, \ldots, p \)), \( E[\delta^2] = 1 \), and \( E[\eta_j^2] = 1 \) (\( j = 0, 1, \ldots, p \)). In addition, we can derive the lower bound as

\[
L = -\frac{n - p - 1}{2} \ln(2\pi) + \frac{2p + 1}{2} - p \ln 2 - \sum_{j=1}^{p} \frac{E[\beta_j^2] E[\eta_j^2]}{2} + \ln \frac{\ln \Var[\beta_j]}{\Gamma(\gamma)} + \ln \Gamma(\gamma) + p \ln \frac{\delta^2}{\Gamma(\delta)} - a_1 \ln b_1 + \ln \Gamma(\delta) + \ln \Gamma(\delta + 1) + \frac{1}{2} \sum_{j=1}^{p} \frac{E[\beta_j^2] E[\eta_j^2]}{2} \gamma^2.
\]

(A21)

where \( \Gamma(\bullet) \) is the gamma function.

Finally, after convergence, we obtain an approximate distribution for each marker effect. The mean \( E[\beta_j] \) can be regarded as a point estimate of the effect of marker \( j \), and its \( 1 - \alpha \) percent credible interval can be defined as

\[
( E[\beta_j] - \sqrt{\Var[\beta_j]Q(\alpha/2)}, E[\beta_j] + \sqrt{\Var[\beta_j]Q(1 - \alpha/2)} )
\]

where \( Q(\bullet) \) here represents the normal quantile function.
Estimation of Quantitative Trait Locus Effects with Epistasis by Variational Bayes Algorithms

Zitong Li and Mikko J. Sillanpää
1 Connection between VB and EM

The VB method we consider in the paper is closely related to the well known EM algorithm. According to TZIKAS et al. (2008), the principle of EM algorithm for the MAP estimation problem $\hat{\theta} = \arg \max_\theta p(\theta|\text{data})$ can be interpreted as follows: First of all, the log posterior distribution can be decomposed as

$$
\ln p(\theta|\text{data}) = \int q(z) \ln \frac{p(\theta, z|\text{data})}{q(z)} dz - \int q(z) \ln \frac{p(z|\theta, \text{data})}{q(z)} dz \\
= L(q(z)) + \text{KL}(q(z)||p(z|\theta, \text{data})),
$$

(1.1)

where $z = \{z_1, ..., z_m\}$ are the hidden variables (to be integrated out), which are not of main interest, but are helpful to estimate the value of parameters $\theta$. $q(z)$ is a free selected distribution for $z$. Since $\text{KL}(q(z)||p(z|\theta, \text{data})) \geq 0$, $L(q(z))$ can be viewed as a lower bound of the log posterior $\ln p(\theta|\text{data})$. Denote the current value of $\theta$ as $\theta^{\text{OLD}}$. In the E-step, the lower bound is maximized with respect to the distribution $q(z)$, which leads to $q(z) = p(z|\theta^{\text{OLD}}, \text{data})$. In the M-step, the lower bound with $q(z)$ substituted by $p(z|\theta^{\text{OLD}}, \text{data})$ is maximized with respect to $\theta$, which leads to a new estimation of $\theta$ as $\theta^{\text{NEW}}$. By implementing these two steps successively, the value of log posterior function increases, and eventually reaches its maximum.

An easy implementation of the EM algorithm requires the full conditional distribution $p(z|\theta, \text{data})$ to be explicitly known. However, in many models, hidden variables may interact with each other, and $p(z|\theta, \text{data})$ becomes too complicated to be used in EM. One solution is to implement the variational approximation method in the E-step, that is, the distribution $q(z)$ is assumed to be of the factorized form $q(z) = \prod_{i=1}^{m} q(z_i)$. Next, the lower bound can be maximized with respect to $q(z_i) (i = 1, 2, ..., m)$. This procedure has been called variational EM (see BEAL 2003). Furthermore, it is possible to consider the parameters $\theta$ as a part of hidden variables, and implement a variational algorithm with only the E-step. This is exactly the same VB method introduced in the paper. Therefore, variational Bayes and variational EM can be viewed as extensions of
traditional EM algorithm, and can be applied to a broader class of Bayesian models.

2 VB estimation algorithm for Bayesian Adaptive Shrinkage

The priors are specified as
\[
p(\beta_0) \propto 1, \quad (2.1)
\]
\[
p(\tau_0^2) \propto \frac{1}{\tau_0^2}, \quad (2.2)
\]
\[
p(\tau_j^2) \propto \frac{1}{\tau_j^2}, \quad (2.3)
\]
\[
p(\beta_j | \tau_j^2) \propto N(\beta_j | 0, \frac{1}{\tau_j^2}), \quad (2.4)
\]
and the logarithm of the joint distribution of parameters and data is
\[
\ln p(\theta, y) = p(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2 | y) \]
\[
= \left[ \frac{n}{2} \ln \tau_0^2 - \frac{\tau_0^2}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \right] + \left[ -\ln \tau_0^2 \right] + \left[ \frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \tau_j^2 \beta_j^2 \right] + \left[ -\sum_{j=1}^{p} \ln \tau_j^2 \right] + C, \quad (2.5)
\]
where C is a constant. Similarly as in VB-EBL, we can derive the VB marginal distributions \(q(\theta | y)\) for parameters \(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2\) in VB-BAS:

(I) **Derivation of** \(q(\beta_0 | y)\): \(q(\beta_0 | y)\) is recognized as a normal distribution with mean
\[
E[\beta_0] = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} E[\beta_j] x_{ij}), \quad (2.6)
\]
and variance
\[
\text{Var}[\beta_0] = \frac{1}{n E[\tau_0^2]}. \quad (2.7)
\]

(II) **Derivation of** \(q(\beta_j | y)\): \(q(\beta_j | y)\) is also a normal distribution, with mean
\[
E[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1} E[\tau_j^2] \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij}, \quad (2.8)
\]
and variance
\[
\text{Var}[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1}, \quad (2.9)
\]
and in addition we need
\[ E[\beta^2_j] = E[\beta_j]^2 + \text{Var}[\beta_j]. \quad (2.10) \]

(III) Derivation of \( q(\tau^2_0|y) \): \( q(\tau^2_0|y) \) is a gamma distribution with parameters

\[
a_1 = \frac{n}{2},
\]

\[
b_1 = \frac{\sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{j=1}^{p} E[\beta_j] x_{ij})^2 + \text{Var}[\beta_0] + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 \text{Var}[\beta_j]}{2}.
\]

and the expectation is
\[ E[\tau^2_0] = \frac{a_1}{b_1}. \quad (2.13) \]

(IV) Derivation of \( q(\tau^2_j|y) \): \( q(\tau^2_j|y) \) is a gamma distribution with parameters

\[
a_{2j} = \frac{1}{2},
\]

\[
b_{2j} = \frac{E[\beta^2_j]}{2},
\]

and the expectation is
\[ E[\tau^2_j] = \frac{a_{2j}}{b_{2j}}. \quad (2.16) \]

Additionally, we initialize the expectation \( E[\beta_j](j = 0, ..., p) \) as 0, and \( E[\tau^2_j](j = 0, ..., p) \) as 2, and then the approximate distributions in (I)-(IV) can be updated successively. Finally, we can derive the lower bound as
\[
L = -\frac{n - 1}{2} \ln(2\pi) + \frac{p + 1}{2} + \ln \text{Var}[\beta_0] + \sum_{j=1}^{p} \ln \text{Var}[\beta_j] - a_1 \ln b_1 + \ln \Gamma(a_1)
\]
\[
- \sum_{j=1}^{p} a_{2j} \ln b_{2j} + \sum_{j=1}^{p} \ln \Gamma(a_{2j}),
\]

where \( \Gamma(\bullet) \) is the gamma function.

3 VB estimation algorithm for Bayesian LASSO

The priors are specified as
\[
p(\beta_0) \propto 1,
\]

\[
p(\tau^2_0) \propto 1/\tau^2_0,
\]

\[
l \sim N(\beta_j | 0, \frac{1}{\tau^2_j}).
\]


\begin{align}
p(\tau_j^2|\lambda^2) & \propto \text{Inv-Gamma}(\tau_j^2|1, \frac{\lambda^2}{2}), \\
p(\lambda^2|\gamma, \nu) & \propto \text{Gamma}(\lambda^2|\gamma, \nu),
\end{align}

and the logarithm of the joint distribution of parameters and data is

\[
\ln p(\theta, y) = p(\beta_0, \beta_1, \ldots, \beta_p, \tau_0^2, \tau_1^2, \ldots, \tau_p^2, \lambda^2|y) \\
= \left[\frac{n}{2} \ln \tau_0^2 - \frac{\lambda^2}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 \right] + \left[\frac{1}{2} \sum_{j=1}^{p} \ln \tau_j^2 - \frac{1}{2} \sum_{j=1}^{p} \tau_j^2 \beta_j^2 \right] + \left[\ln \gamma - \nu \lambda^2 \right] + C,
\]

where \(C\) is a constant. and a VB algorithm is derived as following:

**I** Derivation of \(\hat{q}(\beta_0|y)\): \(\hat{q}(\beta_0|y)\) is recognized as a normal distribution with mean

\[
E[\beta_0] = \frac{1}{n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} E[\beta_j] x_{ij}),
\]

and variance

\[
\text{Var}[\beta_0] = \frac{1}{n E[\tau_0^2]}. 
\]

**II** Derivation of \(\hat{q}(\beta_j|y)\): \(\hat{q}(\beta_j|y)\) is also a normal distribution with mean

\[
E[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1} E[\tau_0^2] \sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{k \neq j} E[\beta_k] x_{ik}) x_{ij},
\]

and variance

\[
\text{Var}[\beta_j] = (E[\tau_0^2] \sum_{i=1}^{n} x_{ij}^2 + E[\tau_j^2])^{-1},
\]

and in addition we need

\[
E[\beta_j^2] = E[\beta_j]^2 + \text{Var}[\beta_j].
\]

**III** Derivation of \(\hat{q}(\tau_0^2|y)\): \(\hat{q}(\tau_0^2|y)\) is a gamma distribution with parameters

\[
a_1 = \frac{n}{2}, \\
b_1 = \frac{\sum_{i=1}^{n} (y_i - E[\beta_0] - \sum_{j=1}^{p} E[\beta_j] x_{ij})^2 + \text{Var}[\beta_0] + \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 \text{Var}[\beta_j]}{2},
\]

and the expectation is

\[
E[\tau_0^2] = \frac{a_1}{b_1}.
\]
IV) Derivation of $\hat{q}(\tau^2_j|y)$: $\hat{q}(\tau^2_j|y)$ is an inverse Gaussian distribution with parameters

$$\mu = \sqrt{\frac{E[\lambda^2]}{E[\beta_j^2]}}$$

(3.15)

and

$$\lambda_0 = E[\lambda^2],$$

(3.16)

and in addition we need the expectation of $\tau^2_j$ as

$$E[\tau^2_j] = \mu,$$

(3.17)

and the expectation of $\frac{1}{\tau^2_j}$ as

$$E\left[\frac{1}{\tau^2_j}\right] = \frac{1}{\mu} + \frac{1}{\lambda_0}.$$  

(3.18)

V) Derivation of $\hat{q}(\lambda^2|y)$: $\hat{q}(\lambda^2|y)$ is a gamma distribution with parameters

$$a_2 = p + \gamma,$$

(3.19)

and

$$b_2 = \frac{1}{2} \sum_{j=1}^{p} E\left[\frac{1}{\tau^2_j}\right] + v,$$

(3.20)

and the expectation of $\lambda^2$

$$E[\lambda^2] = \frac{a_2}{b_2}.$$  

(3.21)

Before running iterative algorithm, we may initialize the expectation $E[\beta_j](j = 0, \ldots, p)$ as 0, $E[\tau^2_j](j = 0, \ldots, p)$ as 2, and $E[\lambda^2]$ as 1. Finally, the lower bound is

$$L = -\frac{n-p-1}{2}\ln(2\pi) + \frac{2p+1}{2} - p\ln 2 + \sum_{j=1}^{p} E[\beta_j^2] E[\tau^2_j] + \frac{\ln \text{Var}[\beta_0]}{2} + \sum_{j=1}^{p} \frac{\ln \text{Var}[\beta_j]}{2}$$

$$-a_1 \ln b_1 + \Gamma(a_1) - \frac{p}{2}\ln \lambda_0 - a_2 \ln b_2 + \Gamma(a_2).$$  

(3.22)

4 MCMC algorithm for Extended Bayesian LASSO

The priors in EBL are specified as

$$p(\beta_0) \propto 1,$$

(4.1)

$$p(\sigma_0^2) \propto 1/\sigma_0^2,$$

(4.2)

$$p(\beta_j | \sigma_j^2) \propto N(\beta_j | 0, \sigma_j^2),$$

(4.3)
\[ p(\sigma_j^2 | \delta^2, \eta_j^2) \propto \text{Exp}(\sigma_j^2 \frac{\delta^2 \eta_j^2}{2}), \]  
\[ p(\delta^2 | \gamma, \nu) \propto \text{Gamma}(\delta^2 | \gamma, \nu), \]  
and
\[ p(\eta_j^2 | \psi, \vartheta) \propto \text{Gamma}(\eta_j^2 | \psi, \vartheta). \]  

A description of the differences between our approach and MUTSHINDA and SILLANPÄÄ (2010) can be seen in the Discussion section. A Gibbs sampler can be built based on the following full conditional posterior distributions for parameters \( \beta_0, \beta_1, \ldots, \beta_p, \sigma_0^2, \sigma_1^2, \ldots, \sigma_p^2, \delta^2, \eta_1^2, \ldots, \eta_p^2 \):

(I) Full conditional posterior distribution of \( \beta_0 \):

\[ p(\beta_0 | \theta, \beta_0, y) \propto N(\beta_0 | \frac{1}{n} \sum_{j=1}^{n} (y_i - \sum_{j}^p x_{ij} \beta_j), \frac{1}{n} \sigma_0^2). \]  

(II) Full conditional posterior distribution of \( \beta_j \):

\[ p(\beta_j | \theta, \beta_j, y) \propto N(\beta_j | \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{k \neq j}^p x_{ik} \beta_k)}{(\sum_{i=1}^{n} x_{ij}^2 + \frac{\sigma_0^2}{\sigma_j^2})}, \frac{\sigma_0^2}{\sum_{i=1}^{n} x_{ij}^2 + \frac{\sigma_0^2}{\sigma_j^2}}). \]  

(III) Full conditional posterior distribution of \( \sigma_0^2 \):

\[ p(\sigma_0^2 | \theta, \sigma_0^2, y) \propto \text{Inv-Gamma}(\sigma_0^2 | \frac{n}{2}, \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j}^p x_{ij} \beta_j)^2}{2}). \]  

(IV) Full conditional posterior distribution of \( \frac{1}{\sigma_j^2} \):

\[ p(\frac{1}{\sigma_j^2} | \theta, \sigma_j^2, y) \propto \text{Inv-Gaussian}(\frac{1}{\sigma_j^2} \sqrt{\frac{\delta^2 \eta_j^2}{b_j^2}}, \delta^2 \eta_j^2). \]  

(V) Full conditional posterior distribution of \( \delta^2 \):

\[ p(\delta^2 | \theta, \delta^2, y) \propto \text{Gamma}(\delta^2 | \gamma, \frac{1}{2} \sum_{j=1}^{p} \eta_j^2 \sigma_j^2 + v). \]  

(VI) Full conditional posterior distribution of \( \eta_j^2 \):

\[ p(\eta_j^2 | \theta, \eta_j^2, y) \propto \text{Gamma}(\eta_j^2 | 1 + \psi, \frac{1}{2} \delta^2 \sigma_j^2 + \vartheta). \]  

In practice, to avoid underflow/overflow, we restrict the value of \( \sigma_j^2 (j = 1, \ldots, p) \) to the interval \([10^{-20}, 10^{20}]\).
Finally, descriptions of the Gibbs sampler for the BAS model can be found in XU (2003), and that for the BL model can be found in YI and XU (2008).

**Literature cited**


MUTSHINDA, C. M., and M. J. SILLANPÄÄ 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics **186**: 1067–1075


YI, N., and S. XU 2008 Bayesian LASSO for quantitative trait loci mapping. Genetics **179**: 1045–1055
File S2

MATLAB Source Code

File S2 is available for download as a compressed folder at
http://www.genetics.org/content/suppl/2011/10/31/genetics.111.134866.DC1.