Theor Appl Genet (2012) 125:1575–1587
DOI 10.1007/s00122-012-1936-1

ORIGINAL PAPER

Swift block-updating EM and pseudo-EM procedures for Bayesian shrinkage analysis of quantitative trait loci

Crispin M. Mutshinda · Mikko J. Sillanpää

Received: 12 April 2011 / Accepted: 3 July 2012 / Published online: 24 July 2012

Abstract
Introduction Virtually all existing expectation-maximization (EM) algorithms for quantitative trait locus (QTL) mapping overlook the covariance structure of genetic effects, even though this information can help enhance the robustness of model-based inferences.
Results Here, we propose fast EM and pseudo-EM-based procedures for Bayesian shrinkage analysis of QTLs, designed to accommodate the posterior covariance structure of genetic effects through a block-updating scheme. That is, updating all genetic effects simultaneously through many cycles of iterations.
Conclusion Simulation results based on computer-generated and real-world marker data demonstrated the ability of our method to swiftly produce sensible results regarding the phenotype-to-genotype association. Our new method provides a robust and remarkably fast alternative to full Bayesian estimation in high-dimensional models where the computational burden associated with Markov chain Monte Carlo simulation is often unwieldy. The R code used to fit the model to the data is provided in the online supplementary material.

Communicated by F. van Eeuwijk.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1936-1) contains supplementary material, which is available to authorized users.

C. M. Mutshinda · M. J. Sillanpää (✉)
Department of Mathematics and Statistics,
University of Helsinki, PO Box 68, 00014 Helsinki, Finland
e-mail: mjs@rolf.helsinki.fi

C. M. Mutshinda
Department of Mathematics and Computer Science,
Mount Allison University, 67 York Street, Sackville,
New Brunswick E4L 1E6, Canada

M. J. Sillanpää
Department of Agricultural Sciences, University of Helsinki,
PO Box 27, 00014 Helsinki, Finland

M. J. Sillanpää
Department of Mathematical Sciences, University of Oulu,
PO Box 3000, 90014 Oulu, Finland

M. J. Sillanpää
Department of Biology, University of Oulu,
PO Box 3000, 90014 Oulu, Finland

Introduction

Identifying the genetic basis (number of genes, along with their effects and genomic positions) of complex phenotypic traits is a fundamental goal of modern genetics. A genomic region that is closely linked to a gene that contributes to the variation in a quantitative trait of interest is called a quantitative trait locus (QTL), and the process of identifying QTLs and evaluating their phenotypic effects is known as QTL mapping.

The mapping of multiple QTLs is typically carried out by regressing the phenotypic trait values of n study individuals on their genotypes at p candidate loci. Here, we focus on experimental crosses derived from inbred lines, more specifically on backcross (BC) or double haploids (DH) progeny, where only two genotypes are possible at any locus (Carbonell et al. 1993; Broman 2001). Letting y_i denote the phenotypic trait value of individual i, we assume that

$$y_i = \mu + \sum_{j=1}^{p} \phi_{ij} b_j + e_i \qquad (1)$$

where ϕ_{ij} is a dummy variable for the genotype of individual i at locus j ($j = 1,\ldots, p$), herein coded as 0 for one genotype and 1 for the other; μ is the intercept; b_j represents the effect of genotype substitution at locus j ($j = 1,\ldots, p$), and e_i ($i = 1,\ldots, n$) are random residual errors assumed to be mutually independent and normally distributed around zero with common variance σ_0^2. Model (1) can be compactly written in matrix form as

$$y = \mu 1_n + \Phi b + e, \tag{2}$$

where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $b = (b_1, \ldots, b_p)^{\mathrm{T}}$, $e = (e_1, \ldots, e_n)^{\mathrm{T}}$, Φ denotes the $n \times p$ design matrix encompassing the genotype profiles of the p loci, and 1_n is the n-dimensional column vector of ones.

In large-scale QTL mapping studies, most of the candidate loci usually have weak or no effect on the quantitative trait of interest (Xu 2003; Yi and Xu 2008; O'Hara and Sillanpää 2009). This is more so when the underlying biology of the trait under study is also sparse. In addition, strong correlations between genotypes of dense markers on the same chromosome induce multicollinearity issues (Xu 2003).

In saturated regression models (i.e., regression models with $p > n$) the ordinary least squares (OLS) method (or equivalently maximum likelihood estimation for linear models with Gaussian residuals) is prone to over-fitting the data, due to a lack of degrees of freedom (curse of dimensionality). The model will essentially adjust to random features of the particular dataset on which it is trained, rather than describing the biologically interesting relationship on which the modeling effort is focused. As a result, the model will achieve a nearly perfect fit to the data at hand while having poor predictive performance (Lande and Thompson 1990; Bishop and Tipping 2003). Multicollinearity issues often exacerbate the difficulty of estimating genetic effects in the presence of fine-scale marker maps, further undermining the usefulness of standard estimation procedures such as the OLS.

When the underlying biology is known to be sparse in large-scale genetic association studies, it becomes essential to seek out a parsimonious or sparse model representation which can adequately describe the genotype-to-phenotype mapping without over-fitting (Kao et al. 1999; Ball 2001; Sen and Churchill 2001; Broman and Speed 2002; Sillanpää and Corander 2002; Yi and Xu 2008). Several methods have been proposed to this end from both the classical and Bayesian perspectives. These can roughly be classified into variable selection and regularization methods (Xu 2007).

Variable selection methods, in the vein of classical stepwise selection techniques Gimelfarb and Lande (1994a, b); Kao et al. 1999; Broman and Speed 2002; Miller 2002) and Bayesian "spike-and-slab" methods like stochastic search variable selection (SSVS; George and McCulloch 1993; Yi et al. 2003; Mutshinda et al. 2009, 2011) and Bayes B-type of methods (Meuwissen et al. 2001; Sillanpää and Bhattacharjee 2005, 2006), involve the idea of pruning (i.e., discarding) the allegedly redundant predictors.

On the other hand, the "selection-free" regularization or shrinkage methods involve all potential predictors, but require (through a suitable penalty function or a sparsity-inducing prior in the Bayesian framework) that spurious effects (i.e., the effects of redundant variables) be automatically shrunken towards zero. Ridge regression (Hoerl and Kennard 1970; Myers 1992; Whittaker et al. 2000; Malo et al. 2008), the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996; Li and Sillanpää 2012b) and their Bayesian analogues (Xu 2003; Wang et al. 2005; Yi and Xu 2008; de los Campos et al. 2009; Sun et al. 2010; Mutshinda and Sillanpää 2010, 2011) fall under the umbrella of shrinkage methods.

A significance effect size threshold for declaring QTLs is typically determined through a permutation-based method (Churchill and Doerge 1994). This method consists of repeatedly fitting the model to the data with the genotypic values held fix while reshuffling the phenotypic values to eliminate causal relationships so that any apparent marker-to-phenotype association can effectively be attributed to chance alone. Retaining the largest (absolute) effect size estimate under each phenotype permutation as test statistic yields an empirical distribution of the test statistic under the null hypothesis of no phenotype-to-genotype association, from which a suitable quantile, e.g., the tenth percentile, can be selected as significance threshold for declaring QTLs.

The Bayesian shrinkage methods for QTL mapping (Xu 2003; Mutshinda and Sillanpää 2010, 2011) and genomic breeding value (GBV) estimation (de los Campos et al. 2009; Cleveland et al. 2010) have so far mostly relied on Markov chain Monte Carlo (MCMC) methods (Gilks et al. 1996; Gelman et al. 2003) for posterior simulation.

However, in the presence of huge amounts of potentially correlated markers, MCMC samplers are prone to poor mixing (slow convergence). Alternative model fitting approaches that can perform fast and yet adequately describe the genotype-to-phenotype relationship without over-fitting are therefore needed. *Maximum a posteriori* (MAP) finding methods are enjoying increasing interest in this respect (Xu 2010; Yi and Banerjee 2009; Sun et al. 2010; Cai et al. 2011; Li and Sillanpää 2012a). It is however often not feasible to analytically maximize the posterior distribution, making the recourse to iterative procedures imperative in many settings.

The expectation-maximization (EM) algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997) provides an effective tool for MAP estimation in many situations. Applications of the EM algorithm in QTL mapping include Xu (2010), Yi and Banerjee (2009) and Sun et al. (2010). Recently, Hayashi and Iwata (2010) adapted the EM algorithm of Yi and Banerjee (2009) for GBV estimation (see also Shepherd et al. 2010; Kärkkäinen and Sillanpää 2012).

Extant EM algorithms for genetic association typically overlook the co-variance structure of genetic effects as these are usually updated one at a time. However, an account for the co-variance information can enhance the model robustness to guarantee that repeated model fitting to the same dataset will produce roughly the same results (Kabán 2007). Most of all, it is more interesting to give a sense of the full posterior distribution of genetic effects, rather than focusing on the MAP point estimates.

Here, we develop fast block-updating EM and pseudo-EM algorithms for Bayesian shrinkage analysis of QTLs designed to provide not only the MAP estimates of genetic effects, but also their posterior covariance matrix involving accuracy estimates (variances) and co-variances that can be useful in subsequent analyses. Here, block-updating implies that all genetic effects are updated in tandem rather than one at a time. We consider two cases here: (1) independence model, where genetic effects are assumed to be independent in their joint prior. For this case, we use a full probability model and our algorithm is a block-updating EM-algorithm. (2) Dependence model, where genetic effects are assumed to be dependent in their joint prior. For this model, we use a pseudo-EM algorithm implying the presence of inconsistency among our updating steps since we are not using a full probability model (Makhuvha et al. 1997; Heckerman et al. 2000; Lunn et al. 2009; Jackson et al. 2009; Lowd and Shamaei 2011). We have done this to simplify computations and speed up the algorithm. The posterior covariance structure of genetic effects obtained from our analysis can be used in Monte Carlo-based predictive inference. We investigate the performance of our method on simulated data, and use it to analyze real data from the North American Barley Genome Mapping project.

Materials and methods

Hierarchical specification of sparsity-inducing priors

A Student's t prior is independently specified on each genetic effect, in a hierarchical fashion. More specifically, we assume that a priori, $b_j|\sigma_j^2 \sim N(0,\ \sigma_j^2)$ and $\sigma_j^2 \sim \mathrm{Inv}-\mathrm{Gamma}(\alpha,\ \lambda)$ or equivalently, $\tau_j = 1/\sigma_j^2 \sim \mathrm{Gamma}\,(\alpha,\ \lambda)$ independently for $(j = 1, \ldots, p)$. Differences in the posteriors, $p(\sigma_j^2|\mathrm{Data})$ of locus-specific variances will induce differential shrinkage of genetic effects across loci.

Along the lines of Yi and Banerjee (2009), we consider the variances σ_j^2 as missing, and integrate them out with the view that the hyper-parameters α and λ can be cautiously selected to induce the desired sparseness property.

The marginal (or unconditional) prior of b_j is obtained as $p(b_j) = \int\limits_0^\infty p(b_j|\tau_j)p(\tau_j)d\tau_j = \lambda^\alpha \Gamma\ \ (\alpha + \frac{1}{2})/(\sqrt{2\pi}\Gamma(\alpha))(\lambda + b_j^2/\ 2)^{-(2\alpha+1)/2}$, where $\Gamma(\cdot)$ denotes Euler's Gamma function: $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\mathrm{d}t,\ z > 0$. It turns out that $p(b_j)$ can be written as

$$p\,(b_j) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sigma\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left\{1 + \frac{1}{v}(b_j/\sigma_j)^2\right\}^{-(v+1)/2}, \qquad (3)$$

and recognized as a Student's t probability density function with scale parameter $\sigma = \sqrt{\lambda/\alpha}$ and $v = 2\,\alpha$ degrees of freedom (for more details see appendix A in Electronic Supplementary Material (ESM), and also Tipping 2001; Tipping and Lawrence 2005).

So, a Student's t prior with specific scale parameter and degrees of freedom can be obtained under this framework by appropriately selecting the hyper-parameters α and λ. In particular, the Jeffreys' prior $p\,(\tau_j) \propto 1/|\tau_j|$ arises when α and λ are set to zero, leading to the improper marginal prior $p(b_j) \propto 1/|b_j|$ with an infinite mode at zero. When v is set to one $(\alpha = 1/2)$, $p(b_j)$ is Cauchy, and $p(b_j)$ becomes $N\,(0,\ \sigma_j^2)$ as $v \to \infty$.

The $\mathrm{Inv}-\mathrm{Gamma}\,(\alpha,\ \lambda)$ prior imposed on the locus-specific variances, σ_j^2, can alternatively be expressed as a scaled inverse chi-square distribution with scale parameter $s^2 = 2\,\lambda/\,v$ and $v = 2\alpha$ degrees of freedom, which we denote as $\mathrm{Inv}-\chi^2(v,\ s^2)$. The probability density function of the $\mathrm{Inv}-\chi^2(v,\ s^2)$ distribution is given by $p\,(\sigma_j^2|v,\ s) \propto (\sigma_j^2)^{-(v+2)/2}\exp\,(-\,v\,s^2/2\,\sigma_j^2)$ (Gelman et al. 2003).

The conjugacy of the Gaussian priors independently assumed on the locus-specific effects, $b_j(j = 1, \ldots, p)$, implies that, conditionally on the variance parameters $\sigma_j^2(j = 1, \ldots, p)$, the vector $b = (b_1, \ldots, b_p)^{\mathrm{T}}$ is a posteriori (multivariate) Gaussian around the MAP estimate $\hat{b}$. An analytical expression of $\hat{b}$ is available in closed form (see below). Moreover, an approximate posterior covariance matrix, $\hat{\Sigma}_b$, of b can be derived in closed form through Laplace's quadratic approximation to the log-posterior around its mode (Gelman et al. 2003).

If we assume uniform priors independently on the residual variance, σ_0^2, and the intercept parameter μ, i.e., $p(\mu,\ \sigma_0^2) \propto 1$, the posterior distribution of σ_0^2 conditionally on b and μ, is Inverse-Gamma with parameters $\alpha = (n/2) - 1$ and $\beta = \frac{1}{2}\sum_{i=1}^n (y_i - \mu - \Phi_i\,b)^2$. Since the mode of the Inverse-Gamma distribution with parameters α and β is given by (see Gelman et al. 2003, p. 574–575) $\beta/(\alpha + 1)$, it follows that the MAP estimate of σ_0^2, conditionally on b and μ, is given by the familiar formula

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu - \Phi_i\, b)^2 \tag{4a}$$

Turning now to the intercept parameter μ, the overall likelihood function based on (2) is $L(y;\ \mu, b,\ \sigma_0^2, \Phi) \propto \exp\{-[(y - \mu\, 1_n\ - \Phi\, b)^{\mathrm{T}}\, \Sigma\, (y - \mu\, 1_n - \Phi\, b)]/2\}$, where $\Sigma = \sigma_0^{-2} I_n$ and I_n is the $n \times n$ identity matrix. Focusing on μ, this likelihood function can be written as $L(y;\ \mu, b,\ \sigma_0^2, \Phi) \propto \exp\{-[(\mu\, 1_n - y - \Phi\, b)^{\mathrm{T}}\, \Sigma\, (\mu\, 1_n - y - \Phi\, b)]/2\}$. Under the uniform prior $p(\mu,\ \sigma_0^2) \propto 1$ assumed here, the conditional posterior of μ given by $p\,(\mu|y,\ b,\ \sigma_0^2)$ is proportional to $\exp\{-[(\mu\, 1_n - y\ - \Phi\, b)^{\mathrm{T}}\, \Sigma\, (\mu\, 1_n - y - \Phi\, b)]/2\}$, which is maximized when $\mu\, 1_n = (y - \Phi\, b)$. So, conditionally on b and σ_0^2, the MAP estimate of μ is given by

$$\hat{\mu} = \mathrm{mean}\,(y - \Phi\, \hat{b})\,. \tag{4b}$$

Before delving into the details of our block-updating EM and pseudo-EM procedures, it is worth giving some insight in the derivation of analytic expressions of $\hat{b}$ and $\hat{\Sigma}_b$, which are the building blocks of our EM and pseudo-EM algorithms.

Analytical expressions of the MAP and the posterior covariance matrix of genetic effects

As already pointed out earlier, the likelihood function for our model based on (2) is $L(y;\ b,\ \sigma_0^2, \Phi) \propto \exp\{-[(y - \mu\, 1_n - \Phi\, b)^{\mathrm{T}}\, \Sigma\, (y - \mu\, 1_n - \Phi\, b)]/2\}$, with $\Sigma = \sigma_0^{-2} I_n$ and I_n denoting the $n \times n$ identity matrix. In the sequel, we drop the conditioning on Φ for ease of notation.

The prior specification for the parameter vector b is $b \sim \mathrm{N}_p(0, \Lambda^{-1})$, where N_p denotes the p-dimensional multivariate normal distribution, $\Lambda = \mathrm{diag}\,(\tau_1, \ldots, \tau_p)$ is a $p \times p$ diagonal matrix comprising effect-specific precisions, $\tau_j = 1/\sigma_j^2$, on the main diagonal. Note that there is a different variance (precision) hyper-parameter for each component of b. This amounts to assigning different weights (corresponding to the idiosyncratic variances) to the columns of Φ, which may result in sparseness as the independent variables corresponding to columns with nearly zero weights are essentially pruned from the model.

The posterior of b results from the combination of the likelihood and prior as

$$\begin{aligned} p(b|y, \mu, \sigma_0^2, \tau)\ \propto\ & \exp\{-[(y - \mu\, 1_n - \Phi\, b)^{\mathrm{T}} \\ & \Sigma\, (y - \mu\, 1_n - \Phi\, b)\ +\ b^{\mathrm{T}} \Lambda\, b]/2\}, \end{aligned} \tag{5}$$

where $\tau = (\tau_1, \ldots, \tau_p)^{\mathrm{T}}$. The MAP or posterior mode of b is given by $\hat{b} = \arg\min_b\{[(y - \mu\, 1_n - \Phi b)^{\mathrm{T}} \Sigma\, (y - \mu\, 1_n - \Phi b)\ + b^{\mathrm{T}} \Lambda\, b]/2\}$, and an approximate posterior covariance matrix of $\hat{b}$ deriving from Laplace's quadratic approximation to the log-posterior around its mode (Gelman et al. 2003) is $\hat{\Sigma}_b = (-\nabla_b \nabla_b \log \mathrm{post}(\hat{b}))^{-1}$. It turns out that $b|y, \sigma_0^2, \tau \sim \mathrm{N}_p(\hat{b},\ \hat{\Sigma}_b)$, with closed-form expressions of $\hat{\Sigma}_b$ and $\hat{b}$ given by (for more details see appendix B in ESM)

$$\hat{\Sigma}_b = (\Phi^{\mathrm{T}} \Sigma\, \Phi + \Lambda)^{-1} = (\sigma_0^{-2}\, \Phi^{\mathrm{T}} \Phi + \Lambda)^{-1}, \tag{6a}$$

$$\hat{b}|\hat{\mu} = \hat{\Sigma}_b\, \Phi^{\mathrm{T}} \Sigma\, (y - \hat{\mu}\, 1_n) = \hat{\sigma}_0^{-2}\, \hat{\Sigma}_b\, \Phi^{\mathrm{T}}\, (y - \hat{\mu}\, 1_n). \tag{6b}$$

Tipping (2001) already derived formulas (6a) and (6b) in a Machine Learning context, by analytically computing the normalizing constant of the posterior distribution $p\,(b|y, \tau, \sigma_0^2) = p\,(y|b,\ \sigma_0^2)\, p\,(b|\tau) \,/ \int p\,(y|b,\ \sigma_0^2)\, p\,(b|\tau)\, d\, b$ as a convolution of two Gaussians. However, it is good to keep in mind that Henderson (1950, 1970) offered essentially these same equations for random effects in the classical Gaussian mixed model context much earlier.

Description of the block-updating EM and pseudo-EM algorithm

Following Yi and Banerjee (2009), we proceed by treating the locus-specific variance (precision) parameters $\sigma_j^2 (\tau_j = 1/\sigma_j^2)$ as missing, and require their conditional expectations given the data and the current estimate of b, i.e., $\hat{\sigma}_j^2 = E\,[\sigma_j^2|y,\ \hat{b}]$ (or equivalently $\hat{\tau}_j = 1/\hat{\sigma}_j^2$) whenever these variance parameters are required.

Given the current estimates $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_p)$, $\hat{\sigma}_0^2$ and $\hat{\mu}$, we know (see above) that $b|y,\ \hat{\mu}, \hat{\tau},\ \hat{\sigma}_0^2 \sim \mathrm{N}_p\,(\hat{b},\ \hat{\Sigma}_b)$, with $\hat{\Sigma}_b$ and $\hat{b}$ defined as in (6a) and (6b). It can also be shown (see appendix C in ESM) that $\sigma_j^2|y,\ \hat{b} \sim \mathrm{Inv}-\chi^2(1 + v_j,\ [v_j\, s_j^2 + \hat{b}_j^2] \,/\, [1 + v_j])$. From the properties of the scaled inverse-chi-square distribution (Gelman et al. 2003, pp. 574–575), it follows that $E[\sigma_j^2|y,\ \hat{b}]\ = (v_j\, s_j^2 + \hat{b}_j^2) \,/\, (v_j - 1)$ for $j = 1, \ldots, p$ and $v_j > 1$.

We are therefore in a situation where the focal parameters, b, have a tractable distribution given latent variables (τ), which in turn have a tractable distribution given b. This enables us to use the EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997) to find $\hat{b}$ and $\hat{\Sigma}_b$.

The EM algorithm alternates between two steps: conditional expectation evaluation (E-step) and maximization (M-step). The E-step of our algorithm consists in replacing each σ_j^2 $(j = 1, \ldots, p)$ by its conditional expectation

$$\hat{\sigma}_j^2 = (v_j s_j^2 + \hat{b}_j^2) / (v_j - 1), \tag{7}$$

$(v_j > 1)$, where $\hat{b}_j$ is the current estimate of b_j or equivalently, replacing each locus-specific precision parameter $\tau_j = 1/\sigma_j^2$ by

$$\hat{\tau}_j = (v_j - 1)/\left(v_j s_j^2 + \hat{b}_j^2\right). \tag{8}$$

For independence model (when prior independence of genetic effects is assumed), the updating steps (7) and (8) follow a full probability model and our algorithm is a standard EM-algorithm. For dependence model (when prior dependence of genetic effects is assumed, see *Smoothing or explicit account of the dependence between genetic effects of nearby loci* below), the updating steps (7) and (8) follow an inconsistent probability model and our algorithm can be called as a pseudo-EM-algorithm. We have done this to simplify calculations.

In the M-step, b is updated through (6a) and (6b), with the estimate $\hat{\Lambda}$ obtained in the E-step, and the current values $\hat{\sigma}_0^2$ and $\hat{\mu}$ of the residual variance and the intercept parameter, respectively, and subsequently updating the last two parameter values through (4a) and (4b).

The E- and M-steps are meant to be iterated until a convergence criterion, (e.g., $||\hat{b}^{(k)} - \hat{b}^{(k-1)}||/||\hat{b}^{(k)}|| < \delta$ for some sufficiently small $\delta > 0$) is satisfied. A sensible choice of initial parameter values is $\sigma_0^2 = \mathrm{var}(y)$, $\mu = \mathrm{mean}(y)$, and all b_j set to be small in absolute value, but $||b|| \neq 0$ is required to avoid a division by zero in the first evaluation of the convergence criterion.

Crucially for us here, the posterior covariance structure of b is accommodated by updating b as a block (cf. ter Braak et al. 2005). However, the implementation of the proposed EM and pseudo-EM algorithms may face computational problems as we discuss next. We also point to a couple of approaches to overcome these issues.

Computational issues

A potential hurdle in implementing our block-updating EM and pseudo-EM algorithms may come from the inversion of the $p \times p$ matrix $(\Phi^T \Sigma\, \Phi + \Lambda)$ involved in (6a), which can be prohibitive when p is very large. In cases where p is much larger than n, and n is relatively small, the Woodbury identity

$$(\Lambda_{p\times p} + \Phi^{\mathrm{T}}_{\mathrm{p\,x\,n}}\, \Sigma_{n\times n}\quad \Phi_{n\times p})^{-1} = \Lambda^{-1}_{p\times p} - \Lambda^{-1}_{p\times p}\, \Phi^{\mathrm{T}}_{p\times n} \\ (\Sigma^{-1}_{n\times n} + \Phi_{n\times p}\quad \Lambda^{-1}_{p\times p}\quad \Phi^{\mathrm{T}}_{p\times n})^{-1}\ \Phi_{n\times p}\quad \Lambda^{-1}_{p\times p} \tag{9}$$

(Zielke 1968; Golub and van Loan 1996; Li et al. 2002) may help overcome this problem by requiring the inverse of the low-dimensional $n \times n$ matrix $(\Sigma^{-1}_{n\times n} + \Phi_{n\times p}\quad \Lambda^{-1}_{p\times p}\quad \Phi^{\mathrm{T}}_{p\times n})$, instead of the larger $p \times p$ matrix $(\Phi^{\mathrm{T}} \Sigma\, \Phi + \Lambda)$, provided that $(\Sigma^{-1}_{n\times n} + \Phi_{n\times p}\quad \Lambda^{-1}_{p\times p}\quad \Phi^{\mathrm{T}}_{p\times n})$ is invertible. The involved $p \times p$ inverse matrix Λ^{-1} is straightforwardly computed by replacing each element on the main diagonal of Λ by its inverse. However, the Woodbury identity may be impractical in large-scale problems with large sample size, n. In such a case, one may require an estimate, $\hat{b}$, of the genetic effects by solving the linear system $(\Phi^{\mathrm{T}} \Sigma\, \Phi + \Lambda)\, b = \Phi^{\mathrm{T}}\, \Sigma\, (\,y - \mu\, 1_n)$ for b through iterative methods such as Gauss-Seidel iteration, in order to avoid the costly matrix inversion involved in (6a). However, while addressing the important issue of scalability, this approach eludes the estimation of the posterior covariance matrix of genetic effects (i.e., $\hat{\Sigma}_b$) and does not provide accuracy estimates. Therefore, using iterative methods to approximate inverse matrix involved in (6a) may be a better alternative in large-scale problems.

Multicollinearity issues arising from intrinsic dependence between marker genotypes at adjacent loci (the columns of Φ) may also induce computational difficulties (associated with matrix inversion). This is more so when dealing with dense marker maps under low recombination rate, as a result of consistent relationships between the genotypes of successive markers. Jittering i.e., adding a tiny random noise to each data point (Gelman and Hill 2007, p. 554) may help prevent the columns of Φ from falling exactly on top of each other, although this is more suitable for continuous variables. Note also that the addition of the diagonal matrix Λ to $\sigma_0^{-2}\, \Phi^{\mathrm{T}}\Phi$ in (6a) may, to some extent, alleviate the collinearity issues associated with the structure of Φ. However, the rescue effect of this operation depends on the magnitude of diagonal values of Λ.

Smoothing or explicit account of the dependence between genetic effects of nearby loci

The covariance matrix involved in our EM algorithm was analytically derived through Laplace's quadratic approximation to the joint posterior of genetic effects. Linkage disequilibrium (LD) induces strong dependency between alleles (genotypes) of nearby loci on the same chromosome. The information about the strength of dependence between the coefficients of neighboring loci can be explicitly incorporated in the model by assuming prior dependence of genetic effects. One way of doing this is to replace the diagonal precision matrix, Λ, involved in (6a) with a full matrix $R = \{r_{i,j}\}$ implementing an exponential decay of the strength of dependence between the genetic effects of loci i and j with the distance (physical or genetic) $d_{i,j}$ separating them. We propose that

$$r_{i,j} = {}^1\!/\!_2\left\{\, \tau_i\, I(i=j)\ + \sqrt{\tau_i\, \tau_j}\ \exp\left(\varphi\, d_{i,j}\right)\right\}, \tag{10}$$

where $I(\cdot)$ denotes the indicator function, and the "smoothing parameter" $\varphi > 0$ is intended to control the degree of dependence between the genetic effects of loci i and j located $d_{i,j}$ units of distance apart from each other on the same chromosome. The value of φ needs to be specified by the analyst in such a way that the ensuing precision matrix $(\sigma^{-2}\Phi^{\mathrm{T}}\Phi + R)$ is invertible. The factor 1/2 in Eq. (10) is introduced to guarantee that the variance parameters in the two versions of the algorithm (without and with smoothing) are on the same scale; in particular, $r_{i,i} = \tau_i$. It is worth keeping in mind that R is (just like Λ) a precision matrix, which explains the absence of a minus sign in the exponent part of (10), since decreasing variance corresponds to increasing precision and vice versa.

The smoothing property enforces the similarity between genetic effects of nearby loci on the same chromosome, which are expected to appear in LD blocks. It may however happen that a locus effect is in opposite direction to the rest of loci of the LD block in which it is supposed to fit. This situation may result from a conflict between the two sources of covariance information involved in $(\sigma^{-2}\Phi^{\mathrm{T}}\Phi + R)$ namely, the data (through $\Phi^{\mathrm{T}}\Phi$) and the prior (through the off-diagonal elements of R). This conflict can, however, be solved by proceeding in two steps. First, fitting the model without smoothing, i.e., updating the genetic effects through Eqs. (6a) and (6b) to identify the directions of locus effects to be used for each LD block, and then applying the reciprocal genotype coding (Conti and Witte 2003; Fridley and Jenkins 2010) to each locus whose effect sign contrasts with that of the LD blocks to which it is supposed to belong.

Finally, it is worth emphasizing that because the Eqs. (7) or (8) used to update the variance components were originally derived assuming independence between loci, the information in the dependence model with smoothing does not go in all directions (Jackson et al. 2009). This means that we are cutting feedback (direct influence) from adjacent genetic effects (due to prior dependence of genetic effects) to the locus-specific variance component. This corresponds to the two-stage estimation approach where in the first stage, variance components are estimated from the model with prior independent loci (without smoothing) and then plugged in the model with prior dependent loci (with smoothing) but with an appropriate account for uncertainty (Heckerman et al. 2000; Lunn et al. 2009; Jackson et al. 2009; Lowd and Shamaei 2011).

Simulation studies

In this section, we report on two simulation studies designed to evaluate the performance of our methodology. In the first simulation study (*Simulation study I*), the analyses are based on the block-updating EM algorithm with (i.e., with no smoothing). The smoothing idea is implemented in *Simulation study II*. Our analyses are based on data with high heritabilities and small sample sizes (which is typical in experimental crosses). However, the methods should perform equally under small heritabilities and large samples as suggested by Sillanpää and Hoti (2007). All computations were performed in R (Development Core Team 2011, http://www.R-project.org) version 2.13.2 on an AMD Turion X2 Dual, equipped with a 64-bit operating system with 2.10 GHz processor and 4 GB of RAM. The R code is provided in the ESM, appendix E.

Simulation study I

In this simulation study, the data generation process was based on two different marker datasets. (1) The moderately dense Barley marker data from the North American Barley Genome Mapping project (Tinker et al. 1996). This dataset involves 145 doubled haploid lines and 127 markers covering seven chromosomes, with an average distance of 10.5 cM between consecutive markers. The original dataset involved 150 DH individuals, but five individuals with missing phenotypes “number of days to heading averaged across 25 different environments” were excluded from the data (this phenotype is considered for real data analysis in the next section). The few missing genotypes were imputed with random draws from Bernoulli (0.5) before the analysis. (2) A dense marker dataset simulated through the WinQTL Cartographer 2.5 program (Wang et al. 2006) and involving 100 BC individuals and 1,000 markers (i.e., 10 times as many markers as individuals) spanning 2 chromosomes, with 500 markers each, and just 1 cM between consecutive markers.

In both cases, the underlying biology was set to be sparse, assuming four QTLs only, namely at loci 4, 25, 50 and 65, with respective genetic effects set to 2.5, −2.5, 4 and −4. In the sequel, loci are identified by their marker indices also referred to as marker numbers. The intercepts were set to zero without loss of generality, and the residual variances were set to 2 and 0.5 in the simulations based on barley marker data and on the synthetic dense marker map, respectively, yielding an average heritability of 0.80 in both cases.

For simulations based on the barley markers, we generated 50 synthetic datasets, and fitted the model to each of them under the following four hyper-parameter settings: $(v = 5,\ s = 0.1)$, $(v = 5,\ s = 0.25)$, $(v = 5,\ s = 0.5)$ and $(v = 2.5,\ s = 0.1)$.

To fit the model to the simulated data based on the dense marker dataset involving 100 BC progeny and 1,000 markers, we used the hyper-parameter settings $(v = 5,\ s = 0.05)$, and utilized the Woodbury identity to handle the large matrix inversion problem.

The convergence was assumed to occur after k iterations if $||\hat{b}^{(k)} - \hat{b}^{(k-1)}||/||\hat{b}^{(k)}|| < 10^{-6}$. In both cases, the algorithm proved to converge after just 20–25 iterations, requiring about 30–45 s for analyses based on the barley marker data, and between 2 and 3 min for the dense marker data.

Figure 1 depicts the posterior means of marker effects averaged over the 50 replicated datasets plotted against the marker indices for simulation based on the barley marker data under the four hyper-parameter settings namely $v = 5$, $s = 0.1$ for panel (a) $v = 5$, $s = 0.25$, for panel (b), $v = 5$, $s = 0.5$ for panel (c), and $v = 2.5$, $s = 0.1$ for panel (d). The broken horizontal lines indicate the permutation-based significance thresholds for declaring QTLs based on 100 phenotype permutations based on a single phenotypic data replicate, and significance level $\alpha = 0.1$.

It is clear from Fig. 1a and d that good separation between actual QTLs and redundant loci can be enforced by setting the hyper-parameter s to be small (typically in the range [0.05, 0.1]), at the cost of missing the covariance structure. The magnitude of noisy signals tends to amplify with increasing s (Fig. 1b, c), thereby increasing the model proneness to false discoveries.

Figure 2a shows a typical plot of the posterior means of genetic effects under the simulated dense marker map, with a suitable hyper-parameter tuning for good separation between QTL and non-QTL loci, namely $(v = 5, \; s = 0.05)$. A zoom in the estimated effects of the first 67 loci (Fig. 2b) reveals that the model-implied QTL positions correspond to the simulated ones. A zoom in the estimated effects of non-QTLs loci, namely over loci 70–300 (Fig. 2c) discloses a positive correlation between the posterior estimates of genetic effects at neighboring loci, which is further corroborated by the posterior autocorrelation function (ACF) or "lagged correlation" of marker effects produced using the R's built-in autocorrelation function acf (R Development Core Team 2011), and displayed in Fig. 2d. In this context, the ACF function at lag k, r_k, indicates the degree of dependence between genetic effects of loci that are k units of distance apart from each other, and is defined by $r_k = \frac{\sum_{i=1}^{p-k}(b_i-\bar{b})\,(b_{i+k}-\bar{b})}{\sum_{i=1}^{p}(b_i-\bar{b})}$ where $\bar{b} = (\sum_{i=1}^{p} b_i)/n$. In time series analysis, the plot of the ACF function as a function of lag is called correlogram. The R's built-in autocorrelation function acf() produces a correlogram along with 95 % confidence bounds for declaring the significance of the autocorrelation at a specific lag $k < p$.

A key feature of our algorithm is its ability to provide accuracy estimates (standard deviations), which can be useful in subsequent inferences. Table 1 gives the posterior modes of genetic effects, along with accuracy estimates in the form of standard deviations derived from the 1,000 × 1,000 posterior covariance matrix, $\hat{\Sigma}_b$, for the four major loci.

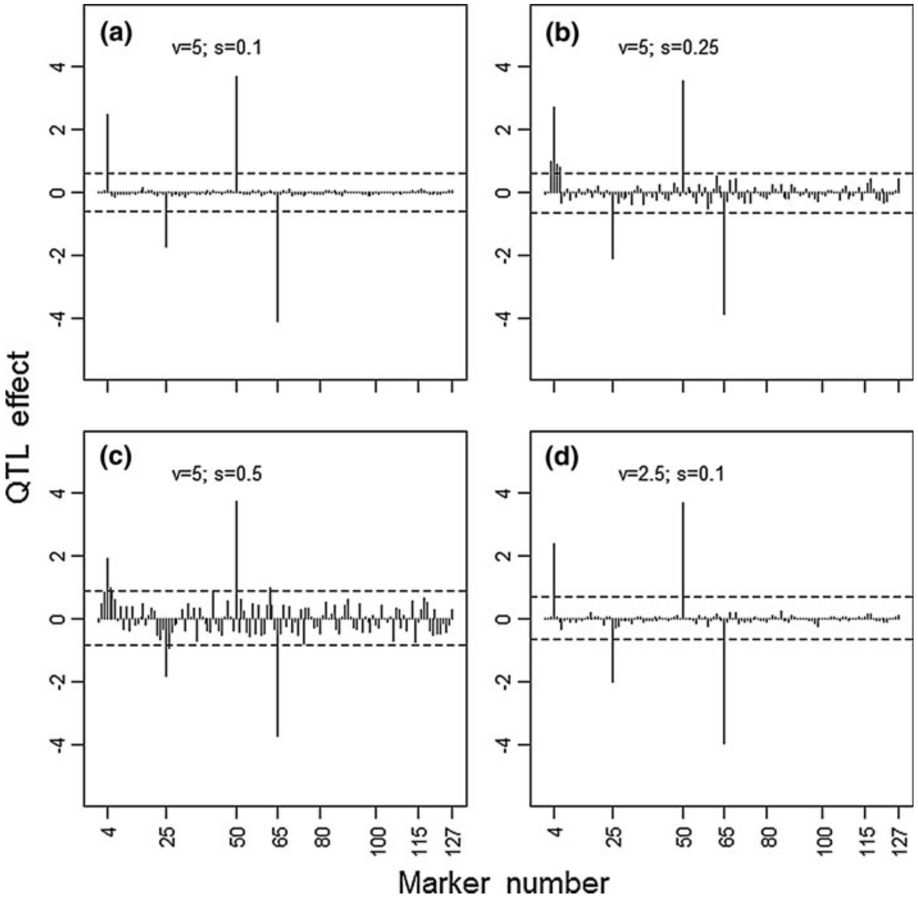

Fig. 1 Posterior modes of QTL effects averaged over 50 replicated datasets plotted against marker numbers under the different hyper-parameter settings shown in the upper part of each panel (**a**–**d**) for simulations based on the barley marker data. The *broken horizontal lines* indicate the permutation-based significance thresholds for declaring QTLs based on 100 phenotype permutations

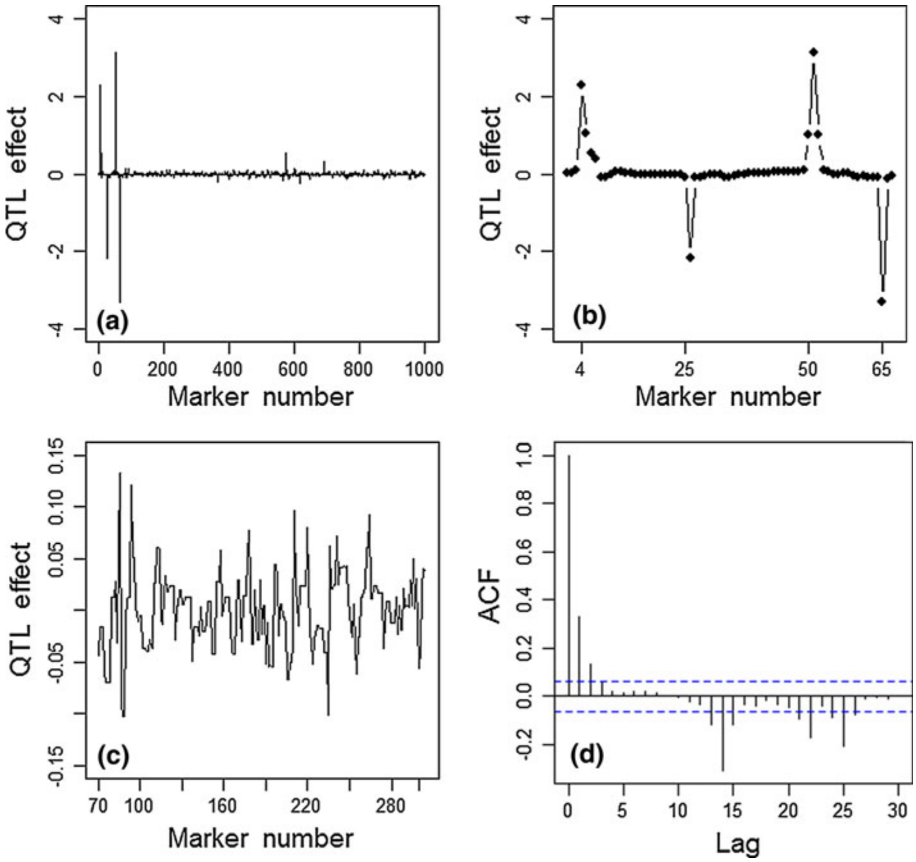

Fig. 2 **a** A typical plot of the posterior modes of genetic effects under the simulated dense marker map, under a suitable hyper-parameter tuning $(v = 5,\ \ s = 0.05)$ for good separation between QTL and non-QTL loci. **b** A zoom in the first 67 markers showing the estimated QTL positions which correspond to the true positions assumed when simulating the data simulation process. **c** A zoom in estimates of allegedly spurious effects over the positions 70–300 illustrating the posterior correlation structure. **d** Posterior autocorrelation function (ACF) of marker effect estimates produced using the *R*'s built-in autocorrelation function acf(). The *dotted horizontal lines* are the 95 % confidence bounds. Significant autocorrelations are indicated by the *vertical lines* crossing the *dashed horizontal ones*

Table 1 True values, posterior modes and accuracy estimates of genetic effects as standard deviations derived from the posterior covariance matrix $\hat{\Sigma}_{\mathbf{b}}$ for the four major loci namely, locus 4, 25, 50 and 65 for the simulated dense marker map

Marker index	True effect	Posterior mode	SD
4	2.5	2.21	0.29
25	−2.5	−2.31	0.10
50	4	3.85	0.37
60	−4	−4.22	0.31

Each locus is identified by its marker index or marker number

Overall, the simulation results corroborate the ability of our modeling approach to identify QTLs, while accounting for the posterior covariance structure of genetic effects.

In Simulation study II below, we illustrate the implementation of the smoothing idea in our block-updating pseudo-EM procedure, and compare the ensuing results to those implied by the "no-smoothing" EM version of the algorithm.

Simulation study II: smoothing

In this simulation study, we considered a synthetic dense marker dataset simulated through the WinQTL Cartographer 2.5 program and involving 100 BC progeny and 500 evenly spaced markers (five times more markers than individuals) over a single chromosome, with 1 cM between consecutive markers. We generated the phenotypic values assuming five QTLs at loci 50, 125, 250, 350 and 480 with respective effects −2.5, 3, −4, −3 and 4. In the data simulation process, the intercept was set to zero and the residual variance to 4, yielding a rough heritability of 0.70. We analyzed the simulated data using our block updating EM and pseudo-EM algorithms (i.e., without and with smoothing). In the latter case, the smoothing parameter φ was set to 10^{-6}. The reported results are based on the hyper-parameter tuning $(v = 5,\ \ s = 0.1)$.

Figure 3 displays a typical plot of the posterior modes of genetic effects against the marker indices for the block updating EM (a) and pseudo-EM (c) algorithms. The corresponding ACFs are plotted in panels (b) and (d), respectively.

As can be seen from Fig. 3, the model was effective at identifying the QTL loci in both no-smoothing and smoothing versions of the procedure. The ACF functions in panels (b) and (b) suggest a clear dependence between the genetic effects of adjacent and nearby loci, the dependency being stronger in the "smoothing version" of the procedure as expected.

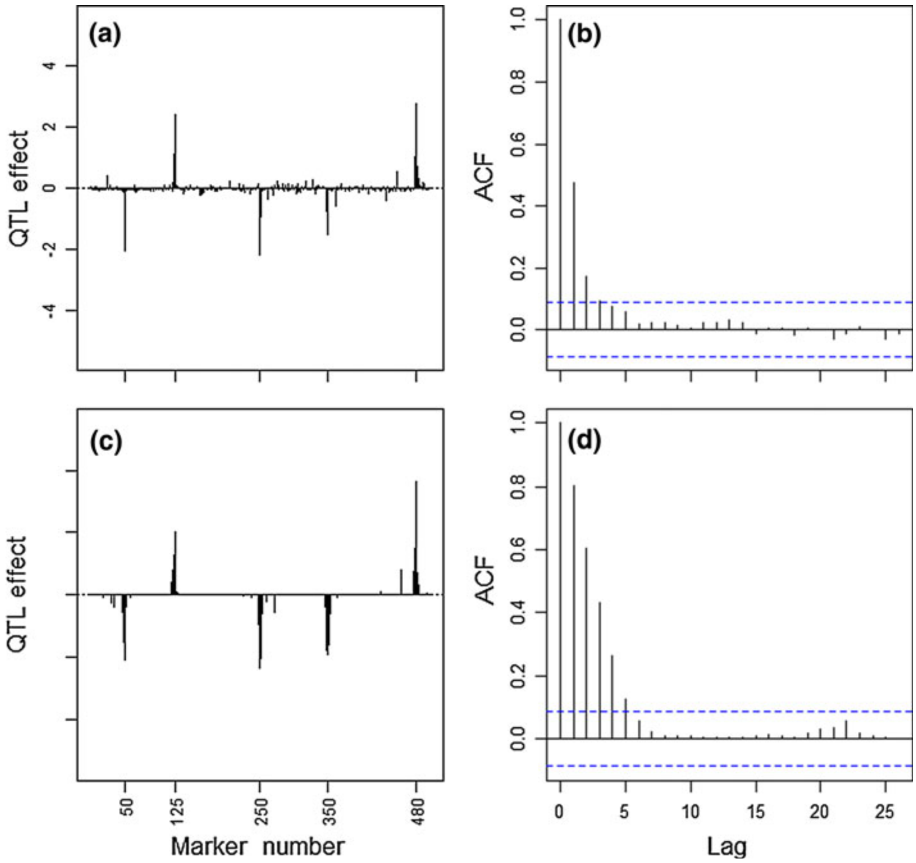

Fig. 3 A typical plot of the posterior modes of genetic effects for the block updating EM procedure **a** without and **c** with smoothing, based on the simulated data involving 500 markers and 100 individuals. The corresponding autocorrelation functions (ACFs) produced using the *R*'s built-in function acf() are shown in panels (**b**) and (**d**), respectively. The *dotted horizontal lines* in panels (**b**) and (**d**) are the 95 % confidence bounds. Significant autocorrelations are indicated by the *vertical lines* crossing the confidence bounds. The simulated QTL loci were 50, 125, 250, 350 and 480 with respective QTL effects −2.5, 3, −4, −3 and 4

Real data analysis

In this section, we use our block-updating EM-based method (without smoothing) to analyze the genetic basis of the time to heading in barley using real-world data from North American Genome Mapping project described in the previous section. The genotypic trait of interest is the number of days to heading averaged over 25 different environments. The data involves 127 markers and 145 doubled haploid lines after 5 individuals with missing phenotype have been omitted as pointed out above. The phenotypic trait values were standardized to have mean zero and unit variance, and the few missing genotypes were imputed with random draws from Bernoulli (0.5) before the analysis. The reported results are based on the hyper-parameter tuning $(v = 5,\ s = 0.05)$.

We used the extended Bayesian LASSO (EBL; Mutshinda and Sillanpää 2010) as benchmark for comparison. We also required Bayes factors (BF; Kass and Raftery 1995; Yi et al. 2007) within SSVS (George and McCulloch 1993; Yi et al. 2003) to evaluate the strength of posterior evidence for including versus not including a particular predictor (locus) in the model. Details on the prior specification for the EBL and SSVS are given in ESM, appendix D. The BUGS code for fitting the EBL to the data is available from Genetics as a supplement to Mutshinda and Sillanpää (2010). The BUGS code for SSVS is provided in ESM, appendix F. The a priori inclusion probability was set to 0.2 (i.e., 0.25 prior odds for inclusion) for all loci.

Here, the BF statistic is nothing but the ratio of the posterior odds to the prior odds for including (H_1) versus not including (H_2) a particular locus in the model. In other words, the BF statistic evaluates the amount by which the prior odds for inclusion of a locus in the model versus its exclusion are changed into posterior odds by the data (i.e., Posterior Odds = Prior Odds × BF). The BF factor for H_1 versus H_2 is often interpreted on the following scale due to Jeffreys (1961). $\mathrm{BF} < 1$: evidence against H_1, $1 < \mathrm{BF} \le 3$: evidence for H_1 no worth than a bare mention, $3 < \mathrm{BF} \le 10$: strong evidence for H_1, $\mathrm{BF} > 10$: decisive evidence for H_1.

For both EBL and SSVS, we ran 20,000 MCMC iterations of two chains, discarding the first 5,000 iterations as burn-in and thinning the remainder to each tenth sample. We assessed the convergence by visually inspecting the mixing of the Markov chains through their traceplots. The 20,000 MCMC iterations took 3,000 s for the EBL and 2,800 s for SSVS.

Figure 4 shows the posterior means of (absolute) genetic effects under the EBL (black circles) and the block-updating EM procedure (grey circles). The dashed horizontal line indicates the permutation-based cutoff effect-size for declaring QTLs based of 100 phenotype permutations under the block-updating EM approach.

The results of the block-updating EM procedure and the EBL were broadly consistent, with loci 6, 9, 12, 33, 40, 47, 63, 86 and 115 emerging as important predictors of the number of days to heading under both models. The SSVS-induced posterior inclusion probabilities for these loci were broadly high, with Bayes factors implying a decisive support for their inclusion in the model.

Recently, Knürr et al. (2011) analyzed the Tinker et al. (1996) data for the same phenotype (time to heading) using two MCMC-based methods namely, SSVS and their newly introduced shrinkage approach based on a mixture of uniform priors. They obtained very similar results (see their Table 1) to those implied by EBL and our swift block-updating EM procedure.

The estimates of the intercept and the residual variances under the block-updating EM were, respectively, $\hat{\mu} = -0.13$ and $\hat{\sigma}_0^2 = 0.14$, the corresponding values under EBL being respectively $\hat{\mu} = -0.11$ and $\hat{\sigma}_0^2 = 0.22$. As expected, the block-updating EM procedure was much faster, providing the results within 1 min. The results of our block-updating EM procedure were robust to the starting values, something which is always at issue in iterative approaches (Xu 2007). This nice feature of our block updating EM procedure is presumably due to the explicit account of the covariance structure of genetic effects.

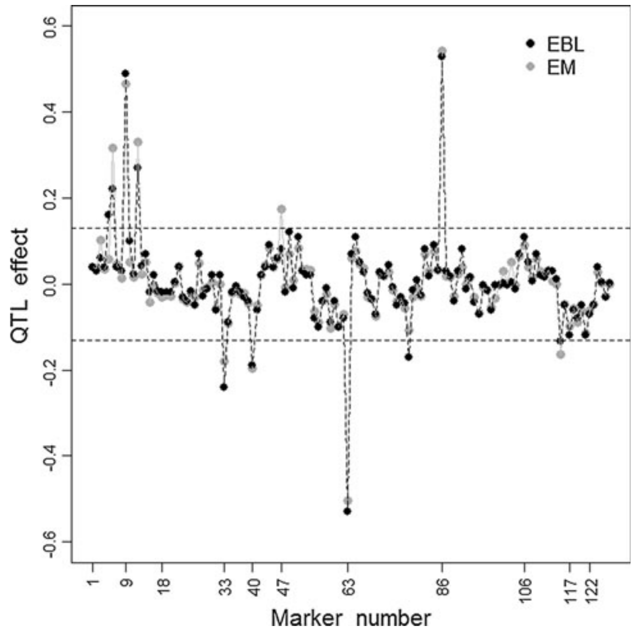

Fig. 4 Posterior means of genetic effects under the MCMC-based EBL (*black circles*) and posterior modes under the block updating EM procedure (*grey circles*) for the barley data, using the number of days to heading as phenotypic trait of interest. The *dashed horizontal line* indicates the permutation-based cutoff effect-size for declaring QTLs based of 100 phenotype permutations under the block-updating EM approach

Discussion

We have proposed fast EM and pseudo-EM-based procedures for Bayesian shrinkage estimation of QTLs, drawing on a block-updating design to accommodate the posterior covariance structure of genetic effects which involves accuracy estimates. In the case where genetics effects are assumed to be dependent in their joint prior, we have called our algorithm pseudo-EM algorithm owing to its similarity with dependency network models where similar Gibbs sampling algorithms are often called pseudo-Gibbs algorithms (Heckerman et al. 2000; Makhuvha et al. 1997). The most well-known application of a pseudo-Gibbs algorithm in genetics is a PHASE haplotyping method (Stephens et al. 2001).

Simulation results demonstrated the effectiveness of our block-updating EM procedure for QTL mapping. This approach is particularly suitable in the presence of highly correlated loci as methods that independently update the model effects tend to select only one locus from a group of highly correlated ones (Zou and Hastie 2005), with the tendency to select different loci in different model runs with the same data. The ability to accommodate the posterior covariance structure of genetic effects keeps our block-updating approach afar from this flaw. It has been suggested (Kabán 2007) that a proper account of the covariance structure helps enhance the robustness of the model along with its predictive performance.

We discussed the idea of smoothing intended to enforce the dependency between genetic effects of nearby loci on the same chromosome through the prior specification, for which the pseudo-EM algorithm is appropriate, and demonstrated its implementation with simulated inbred line cross data. The smoothing approach may be even more useful in outbred populations where the LD may arise from many different sources (Conti and Witte 2003; Sillanpää and Bhattacharjee 2005).

We used our new block-updating EM procedure to analyze the genetic architecture of the number of days to heading in Barley, using data from North American Genome Mapping project (Tinker et al. 1996). We also fitted the extended Bayesian LASSO (EBL; Mutshinda and Sillanpää 2010) to the same data for the sake of comparison, and required SSVS-induced Bayes factors to evaluate the strength of posterior evidence for including each locus in the model.

The results were broadly consistent between the block-updating EM-based approach and the MCMC-based EBL (Fig. 4), with the former being by far faster. In both cases, loci 6, 9, 12, 33, 40, 47, 63, 86 and 115 appeared to contribute to the variation in the time to heading. The SSVS-induced Bayes factors for the inclusion versus exclusion of each of these loci were large enough to imply a decisive support for their inclusion in the model according to the Jeffreys' scale (1961).

Yi and Banerjee (2009) proposed an apparently block-updating EM algorithm for QTL mapping based on the iterated weighted least squares (IWLS) method, with a diagonal weighting matrix involving a different variance parameter for each marker effect as required for differential shrinkage across loci. However, their method treats the regression coefficients independently before a new iteration, by contrast to the methodology proposed here, where the covariance structure is meant to be updated alongside the regression parameters.

The posterior covariance structure of genetic effects can be easily incorporated into a Monte Carlo (MC) analysis involving functions of genetic effects such as breeding values. The MC analysis in this case consists in repeatedly simulating from the approximate joint posterior of genetic effects, and evaluating the quantity of interest. This yields an empirical distribution of the quantity of interest upon which statistical conclusions can be based. For example, the probability for a genetic effect to exceed a specific threshold can be evaluated by the proportion of simulated samples where the effect exceeds the threshold (cf. Hoti and Sillanpää 2006).

The tuning of hyper-parameter remains a critical issue for the performance of EM-based algorithms, in particular in the QTL mapping context as pointed out by Yi and Banerjee (2009) and Xu (2010). There is no trivial solution to this problem, since the most suitable hyper-parameter values are typically data-dependent. In our new approach, the hyper-parameters can be duly tuned to enforce either a clear separation between QTL and non-QTL loci by setting the hyper-parameter s to be small (typically a value in the range 0.05–0.1). The dependence between genetic effects at nearby loci can also be enforced through the pseudo-EM version. Whilst good separation is the point of QTL mapping, an account for the posterior covariance structure of genetic effects is important for enhanced phenotype prediction.

As a cautionary remark, the covariance structure in focus here refers to associations between genetic effects in their joint distribution, which is not to be confused with the interaction effects involving two or more genes, known as epistasis (Carlborg and Andersson 2002; Xu and Jia 2007). The block-updating EM algorithm (the no-smoothing version) can be tailored for epistatic search with $\Lambda \in \Re^{q \times q}$, where $q = (p^2 + p)/2$ is the total number of main and pairwise epistatic effects (cf. Li and Sillanpää 2012a). This may provide a better alternative to the single-component updating EM algorithm for epistatic search, but this requires empirical assessment.

In conclusion, the block-updating EM and pseudo-EM methods proposed in this paper provide expedient alternatives to full Bayesian estimation in high-dimensional models where the computational challenges associated with MCMC are typically awkward.

Acknowledgments The authors wish to thank Hanni Kärkkäinen, Zitong Li and two anonymous referees for their pertinent comments and suggestions. This work was supported by a research grants from the Academy of Finland and University of Helsinki's research funds.

References

Ball RD (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using Bayesian information criterion. Genetics 159:1351–1364

Bishop CM, Tipping ME (2003) Bayesian regression and classification. In: Suykens J, Horvath G, Basu S, Micchelli C, Vandewalle J (eds) Advances in learning theory: methods, models and applications, vol 190. IOS Press, NATO Science, Amsterdam, pp 267–285

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J Roy Stat Soc B 64:641–656

Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Anim 30:44–52

Cai X, Huang A, Xu S (2011) Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinform 12:211

Carbonell EA, Asins MJ, Baselga M, Balansard E, Gerig TM (1993) Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid populations. Theor Appl Genet 86:411–416

Carlborg Ö, Andersson L (2002) Use of randomization testing to detect multiple epistatic QTLs. Genet Sel Evol 79:175–184

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Cleveland MA, Forni S, Nader D, Maltecca C (2010) Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. BMC Proc 4(Suppl 1):S6

Conti DV, Witte J (2003) Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. Am J Hum Genet 72:351–363

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–38

Fridley BL, Jenkins GD (2010) Localizing putative markers in genetic association studies by incorporating linkage disequilibrium into Bayesian hierarchical models. Hum Hered 70:63–73

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman and Hall, New York

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88:881–889

Gilks WR, Richardson S, Spiegelhalter DJ (eds) (1996) Markov Chain Monte Carlo in practice. Chapman and Hall, London

Gimelfarb A, Lande R (1994a) Simulation of marker-assisted selection in hybrid populations. Genet Res 63:39–47

Gimelfarb A, Lande R (1994b) Simulation of marker-assisted selection for non-additive traits. Genet Res 64:127–136

Golub G, van Loan C (1996) Matrix computations, 3rd edn. The John Hopkins University Press, Baltimore

Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet 11:3

Heckerman D, Chickering DM, Meek C, Rounthwaite R, Kadie C (2000) Dependency network for inference, collaborative filtering, and data visualization. J Mach Learn Res 1:49–75

Henderson CR (1950) Estimation of genetic parameters. Ann Math Stat 21:309–310

Henderson CR (1970) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Hoti F, Sillanpää MJ (2006) Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. Heredity 97:4–18

Jackson CH, Best NG, Richardson S (2009) Bayesian graphical models for regression on multiple data sets with different variables. Biostatistics 10:335–351

Jeffreys H (1961) Theory of probability. Clarendon Press, Oxford

Kabán A (2007) On Bayesian classification with Laplace priors. Patt Rec Lett 28:1271–1282

Kao C-H, Zeng Z-B, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Knürr T, Läärä E, Sillanpää MJ (2011) Genetic analysis of complex traits via Bayesian variable selection: the utility of a mixture of uniform priors. Genet Res 93:303–318

Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191:969–987

Lande R, Thompson R (1990) Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics 124:743–756

Li Y, Campbell C, Tipping ME (2002) Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18:1332–1339

Li Z, Sillanpää MJ (2012a) Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231–249

Li Z, Sillanpää MJ (2012b) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419–435

Lowd D, Shamaei A (2011) Mean field inference in dependency networks: an empirical study. In: Proceedings of the 25th conference on artificial intelligence (AAAI-11), San Francisco, CA

Lunn D, Best N, Spiegelhalter D, Graham G, Neuenschwander B (2009) Combining MCMC with 'sequential' PKPD modelling. J Pharmacokinet Pharmacodyn 36:19–38

Makhuvha T, Pegram G, Sparks R, Zucchini W (1997) Patching rainfall data using regression methods. 1. Best subset selection, EM and pseudo-EM methods: theory. J Hydrol 198:289–307

Malo N, Libiger O, Schork NJ (2008) Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet 82:375–385

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Mutshinda CM, O'Hara RB, Woiwod IP (2011) A multispecies perspective on ecological impacts of climatic forcing. J Anim Ecol 80:101–107

Mutshinda CM, Sillanpää MJ (2011) Bayesian shrinkage analysis of QTLs under shape-adaptive shrinkage priors, and accurate re-estimation of genetic effects. Heredity 107:405–412

Mutshinda CM, Sillanpää MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186:1067–1075

Mutshinda CM, O'Hara RB, Woiwod IP (2009) What drives community dynamics? Proc R Soc B 276:2923–2929

Miller A (2002) Subset selection in regression. Chapman and Hall, London

Myers RL (1992) Classical and modern regression analysis, 2nd edn. Wiley, New-York

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian Anal 4:85–118

R Development Core Team (2011) R: A language and environment for statistical computing, reference index version 2.13.2. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387

Shepherd R, Meuwissen THE, Woolliams JA (2010) Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. BMC Bioinform 11:529

Sillanpää MJ, Hoti F (2007) Mapping quantitative trait loci from a single tail sample of the phenotype distribution including survival data. Genetics 177:2361–2377

Sillanpää MJ, Bhattacharjee M (2006) Association mapping of complex trait loci with context-dependent effects and unknown context-variable. Genetics 174:1597–1611

Sillanpää MJ, Bhattacharjee M (2005) Bayesian association-based fine mapping in small chromosomal segments. Genetics 169:427–439

Sillanpää MJ, Corander J (2002) Model choice in gene mapping: what and why. Trends Genet 18:301–307

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Sun W, Ibrahim JG, Zou F (2010) Genome-wide multiple loci mapping in experimental crosses by the iterative penalized regression. Genetics 185:349–359

ter Braak CJF, Boer MP, Bink MCAM (2005) Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170:1435–1438

Tibshirani R (1996) Regression shrinkage and selection via LASSO. J Roy Stat Soc B 58:267–288

Tinker NA, Mather DE, Rosnagel BG, Kasha KJ, Kleinhofs A (1996) Regions of the genome that affect agronomic performance in two-row barley. Crop Sci 36:1053–1062

Tipping ME, Lawrence ND (2005) Variational inference for Student-t models: robust Bayesian interpolation and generalized component analysis. NeuroComputing 69:123–141

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1:211–244

Wang S, Basten CJ, Zeng Z-B (2006) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC

Wang H, Zhang Y-M, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170:465–480

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Xu S (2010) An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 105:483–494

Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63:513–521

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789–801

Xu S, Jia Z (2007) Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics 175:1955–1963

Yi N, Banerjee S (2009) Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181: 1101–1113

Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179:1045–1055

Yi N, George V, Allison DB (2003) Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 164:1129–1138

Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS (2007) An efficient Bayes model selection approach for interacting quantitative trait loci models with many effects. Genetics 176:1865–1877

Zielke G (1968) Inversion of modified symmetric matrices. J Assoc Comput Mach 15:402–408

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc B 67:301–320