Bayesian Multilocus Association Models for Prediction and Mapping of Genome-Wide Data

DOCTORAL THESIS IN ANIMAL SCIENCE

Hanni P. Kärkkäinen

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Agriculture and Forestry of the University of Helsinki, for public criticism in the Lecture Hall of Koetilantie 5, Helsinki, on November 15th 2013, at 12 o'clock noon.

Helsinki 2013

DEPARTMENT OF AGRICULTURAL SCIENCES | PUBLICATIONS | 20

Supervisor: **Professor Mikko J. Sillanpää**
University of Oulu
Department of Mathematical Sciences
Department of Biology and Biocenter Oulu
P.O.Box 3000
FIN – 90014 Oulu, Finland

Co-supervisor: **Adjunct Professor Jarmo Juga**
University of Helsinki
Department of Agricultural Sciences
P.O.Box 27
FIN – 00014 Helsinki, Finland

Reviewers: **Professor Daniel Sorensen**
Aarhus University
Department of Molecular Biology and Genetics
P.O.Box 50
DK – 8830 Tjele, Denmark

Professor Otso Ovaskainen
University of Helsinki
Department of Biosciences
P.O. Box 56
FIN – 00014 Helsinki, Finland

Opponent: **Senior Researcher Luc Janss**
Aarhus University
Department of Molecular Biology and Genetics
P.O.Box 50
DK – 8830 Tjele, Denmark

ISBN 978-952-10-8873-5 (Paperback)
ISBN 978-952-10-8874-2 (PDF)
Electronic publication at http://ethesis.helsinki.fi
Unigrafia
Helsinki 2013

List of original publications

The following original papers are referred in the text by their Roman numerals.

(I) Kärkkäinen, H. P. and M. J. Sillanpää, 2012 Back to basics for Bayesian model building in genomic selection. Genetics 191:969–987.

(II) Kärkkäinen, H. P. and M. J. Sillanpää, 2012 Robustness of Bayesian multilocus association models to cryptic relatedness. Ann. Hum. Genet. 76:510–523. Corrected by: Corrigendum. Ann. Hum. Genet. 77:275.

(III) Kärkkäinen, H. P. and M. J. Sillanpää, 2013 Fast genomic predictions via Bayesian G-BLUP and multilocus models of threshold traits including censored Gaussian data. G3 (Bethesda) 3:1511–1523.

The publications have been reprinted with the kind permission of their copyright holders.

The contributions of the authors HPK and MJS can be detailed as follows:

I Both authors were involved in the conception and design of the study. HPK derived the fully conditional posterior distributions and the GEM algorithm, implemented the algorithm with Matlab, performed the data analyses and drafted the manuscript. Both authors participated in the interpretation of results and critically revised the manuscript.

II Both authors were involved in the conception and design of the study. HPK derived the fully conditional posterior distributions and the GEM algorithm, implemented the algorithm with Matlab, performed the data analyses and drafted the manuscript. Both authors participated in the interpretation of results and critically revised the manuscript.

III Both authors were involved in the conception and design of the study. HPK derived the fully conditional posterior distributions and the GEM algorithm, implemented the algorithm with Matlab, performed the data analyses and drafted the manuscript. Both authors participated in the interpretation of results and critically revised the manuscript.

Contents

Foreword

Genome-wide marker data is used in animal and plant breeding in computing genomic breeding values, and in human genetics in identifying disease susceptibility genes, predicting unobserved phenotypes and assessing disease risks. While the tremendous number of markers available for easy and cost-effective genotyping is an invaluable asset in genetic research and animal and plant breeding, the ever increasing data sets are placing heavy demands on the statistical analysis methodology. The statistical methods proposed for genomic selection are based on either traditional best linear unbiased prediction (BLUP) or different Bayesian multilocus association models. In human genetics the most prevalent approach is a single SNP association model. The thesis consists of three original articles trying to obtain further understanding of the behavior of the different Bayesian multilocus association models and of the instances in which different methods work best, to seek connections between the different Bayesian models, and to develop a Bayesian multilocus association model framework, along with an efficient parameter estimation machinery, that can be utilized in phenotype prediction, genomic breeding value estimation and quantitative trait locus (QTL) location and effect estimation from a variety of genome-wide data.

1 Introduction

The invention of single nucleotide polymorphisms (SNP) in conjunction with the utilization of microarray technology in high-throughput genotyping has exploded the availability of genome-wide sets of molecular markers. Whole genome SNP chips are available for a wide range of species, including humans, agriculturally important plant and animal species, and genetic model organisms. In human genetics the common goal of a genome-wide association (GWA) study is to detect disease susceptibility genes, predict unobserved phenotypes, and assess disease risks at the individual level (Lee *et al.* 2008; de los Campos *et al.* 2010). The animal and plant breeders, on the other hand, are mainly interested in estimating genomic breeding values for genomic selection (Eggen 2012; Nakaya and Isobe 2012).

Genomic selection refers to marker assisted selection using a genome-wide marker information directly in predicting genomic breeding values, rather than first identifying the causal genes (Meuwissen *et al.* 2001). The basic principle of genomic selection includes a set of individuals, known as the training set or the reference population, with phenotypic records and genotypic information of a whole-genome SNP array, and a statistical model explaining the connection between the marker genotypes and the phenotypic observations. The training set data is employed in estimating the effects of the SNP markers or genotypes to the phenotype, that is, the parameters of the model. The acquired information is then used in predicting the heritable part of the phenotype, *i.e.* genomic breeding value, of new individuals (the prediction set) that have only genotypic information available. In animal and plant breeding, the most commonly used approach to predict genomic breeding values based on molecular markers is the genomic best linear unbiased prediction or G-BLUP, a direct descendant of the pedigree-based best linear unbiased prediction (BLUP) model (Henderson 1975). G-BLUP employs the marker information in estimating genomic relationships between the individuals, and utilizes the marker-estimated genomic relationship matrix in a mixed model context (*e.g.* VanRaden 2008; Powell *et al.* 2010). A relatively recent but promising contender for the BLUP-type of model in the genomic selection field is to apply simultaneous estimation and variable selection or regularization to multilocus association models (*e.g.* Meuwissen *et al.* 2001; Xu 2003). A multilocus association model uses the marker information directly by assigning different, possibly zero, effects to the marker alleles and quantifies the genomic breeding value of an individual as the sum of the marker effects. The advantage of a multilocus association over G-BLUP is that the former allows the estimated effect size to vary over the set of markers, while the latter assumes a constant impact throughout the genome.

In human genetics the genome-wide association methods are mainly used for mapping of complex genetic traits. Association mapping utilizes the linkage disequilibrium (LD) between the markers and the causal loci in locating the actual causal genes by searching associations between the markers and the phenotype. Population-based association analyses are more powerful than within-family analyses in detecting the genetic loci associated with the phenotype of interest. As a draw-back, the population-based studies often suffer from an inflated rate of false positives due to population stratification (*i.e.* model misspecification in the presence of hidden population structure) and cryptic relatedness (*i.e.* model misspecification in the presence of sample structure) (see Kang *et al.* 2010). For example, if two populations in Hardy-Weinberg proportions with divergent allele frequencies are combined, the combined population may have large amount of linkage disequilibrium simply due to the combination (*e.g.* Ewans and Spielman 1995). Equivalently, the sample structure of the data may lead to allelic association caused by close relatedness between the individuals rather than true association between the marker and the trait. As *e.g.* PLINK (Purcell *et al.* 2007) omits the sample and population structure from the model, the artificial linkage disequilibrium is likely to cause false positive and negative signals for marker loci without any connection to the studied trait. Although some other heavily-used association methods, including *e.g.* TASSEL (Bradbury *et al.* 2007), GenABEL (Aulchenko *et al.* 2007), EMMA (Kang *et al.* 2008) and EMMAX (Kang *et al.* 2010), provide a sample structure correction, they consider only one marker at the time, ignoring the possible effects of the other major loci. This is less than ideal in genome-wide study for a complex trait, as such traits are assumed to be affected by a multitude of genes (Weeks and Lathrop 1995).

The problem with a multilocus association model applied to a genome-wide data set is oversaturation: since usually the number of SNP markers is orders of magnitude greater than the number of individuals, there are far more explanatory variables than observations in the model. This leads to a situation where some kind of selection or regularization of the predictors is required, either by selecting a subset of the variables that explains a large proportion of the variation, by using orthogonal or nonorthogonal combinations of the variables, or by shrinking the effects of the variables towards zero (*e.g.* Sillanpää and Bhattacharjee 2005; Hoggart *et al.* 2008; O'Hara and Sillanpää 2009; Wu *et al.* 2009; Ayers and Cordell 2010; Cho *et al.* 2010). The appeal in the shrunken estimates is that these methods keep the dimension constant across the possible models by not actually selecting a subset of variables, but instead setting the effect of unimportant ones to (or near) zero. The drawback is that the estimates tend to be biased towards too small values. The methods discarding markers irrelevant to the phenotype are often referred as variable selection, while the ones assigning a penalty term to shrink the marker effects towards zero are considered as variable regularization.

Contrary to the frequentist way of deriving a shrinkage estimator by subtracting a penalty from the gain function (in other words, by adding a penalty to the loss function), in the Bayesian context the regularization mechanism is included into the model by specifying an appropriate prior density for the regression coefficients. A penalized maximum likelihood estimate for the regression coefficients $\boldsymbol{\beta}$ is acquired by maximizing the penalized gain function

$$\widehat{\boldsymbol{\beta}}_{\text{PML}} = \arg\max_{\boldsymbol{\beta}} \;\; \log(p(data|\boldsymbol{\beta})) - \lambda J(\boldsymbol{\beta}), \tag{1.1}$$

where $\log(p(data|\boldsymbol{\beta}))$ is the log likelihood and $J(\boldsymbol{\beta})$ a penalty function. Commonly used penalty functions are derived from the sum of the L2 or L1 norms of the regression coefficients,

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\beta_j||_2^2 = \sum_{j=1}^{p} \beta_j^2 \quad \text{and} \quad J(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\beta_j||_1 = \sum_{j=1}^{p} |\beta_j|,$$

leading to Ridge Regression (Hoerl 1962) and LASSO (Tibshirani 1996) estimates, respectively. The frequentist penalty function is connected to the prior density of a Bayesian model, as the exponent of the function maximized in the frequentist method equals the product

$$\exp(\,\log(p(data|\boldsymbol{\beta})) - \lambda J(\boldsymbol{\beta})) = p(data|\boldsymbol{\beta})\exp(-\lambda J(\boldsymbol{\beta})), \tag{1.2}$$

where $p(data|\boldsymbol{\beta})$ is the likelihood and $\exp(-\lambda J(\boldsymbol{\beta}))$ represents the prior density function. For example, it can be easily seen that the Ridge Regression penalty equals a Gaussian prior density, as $\exp(-(1/\lambda)\sum_{j=1}^{p}\beta_j^2)$ is a kernel of a Gaussian probability density function. Similarly the L1 penalty equals a double exponential or Laplace density. Although it is clearly more logical to consider the assumptions about the model sparseness as a part of the model (the prior is a part of the model) rather than a part of the estimator (a penalty is a part of the estimator), the difference may seem trivial in practice. However, the fact that in Bayesian context the model includes all available information, permits the estimator to be always the same, either the whole posterior density or a *maximum a posteriori* (MAP) point estimate, which in turn enables a straightforward translation of the model into an algorithm.

In the Bayesian context the variable regularization is included into the model by specifying a "spike and slab" prior for the regression coefficients, with "spike" being the probability mass centered near zero and "slab" the probability mass distributed over the nonzero values (see O'Hara and Sillanpää 2009). This prior represents the assumption that only a small proportion of the predictors have a non-negligible effect ("slab"), while the majority of the effects are close to zero ("spike").

The Bayesian models proposed in the literature differ with respect to the "spike and slab" prior densities given for the regression coefficients. The desired shape for the prior density may be acquired either as a mixture of two densities, in which case the model includes a dummy variable indicating whether the effect of a given explanatory variable comes from the "spike" or from the "slab" part of the prior, or alternatively a single prior density approximating the "spike and slab" -shape may be assigned directly on the regression coefficients. In the latter case, the probability density functions commonly used for imitating the "spike and slab" shape are Student's t (*e.g.* Bayes A by Meuwissen *et al.* 2001; Xu 2003; Yi and Banerjee 2009) and Laplace densities (*e.g.* Park and Casella 2008; Yi and Xu 2008; de los Campos *et al.* 2009; Xu 2010; Li *et al.* 2011). Due to the connection to the frequentist L1 penalty function the models with a Laplace prior density are commonly denoted as Bayesian LASSO (Park and Casella 2008). Both Student's t and Laplace density functions possess several favorable features, including high kurtosis and heavy tails, that make them worthy candidates for shrinkage inducing priors. Compared to Gaussian density, these functions consist of a greater probability mass centered near zero and higher probability for large values inducing strong shrinkage to the intermediate sized estimate values and proportionally less shrinkage to the large values and the values near zero. While a Gaussian prior density, or equivalently frequentist Ridge Regression, assigns same penalty to all of the regression coefficients, the heavy-tailed functions work by producing a clearer distinction between large and small estimate values by pushing the intermediate sized values to either direction. For this reason the method is sometimes denoted as adaptive shrinkage.

Several modifications of the indicator-type methods have been introduced, differing with respect to the mixture components (distributions that are used to form the mixture distribution) set for the regression coefficients and the hierarchical structure of the prior (the dependency between the indicator and the marker effect, and the participation of the indicator in the likelihood). While the stochastic search variable selection (SSVS) models considers the "spike and slab" as a mixture of two normal distributions (George and McCulloch 1993; Verbyla *et al.* 2009), or two Student's t distributions (*e.g.* Yi *et al.* 2003), majority of the methods straightforwardly set the regression coefficient to be zero when the indicator is zero (so the "spike" is in fact a point mass located at zero). A prior consisting a mixture of a Student's t density and a point mass at zero has been used in several methods, including BayesB (Meuwissen *et al.* 2001), Hayashi and Iwata (2010) and Habier *et al.* (2011). A similar mixture based on a Laplace density has been used by Meuwissen *et al.* (2009) and Shepherd *et al.* (2010). The simplest hierarchical structure of the prior density, proposed by Kuo and Mallick (1998), determines the effect of the marker j to the phenotype as a product of the indicator γ_j and the effect size β_j, and considers these two to be *a priori* independent. Hence the joint prior of the marker effect $\gamma_j \beta_j$ becomes simply $p(\gamma_j \beta_j) = p(\gamma_j)p(\beta_j)$, where $p(\gamma_j)$ is a Bernoulli density with a prior probability for a marker to be linked to the trait and $p(\beta_j)$ is the Gaussian, Student's t or Laplace prior density given for the effect size. Other types of hierarchical structures presented in the literature include BayesB (Meuwissen *et al.* 2001) where the marker effect is given by β_j alone since the likelihood does not include the indicator; instead, the indicator acts through the effect variance. In Gibbs variable selection, on the other hand, the marker effect is considered as a product of the indicator and the effect size, but the prior density of the effect size is dependent on the indicator (Dellaportas *et al.* 2002).

Whether the model is based on a Student's t, Laplace, or a mixture prior density, the intensity of the shrinkage produced by the prior is determined by the prior parameters (*i.e.* hyperparameters) defining the shape of the prior density function. The models proposed in the literature differ from each other in terms of the procedures they use to determine the prior parameters. In the original BayesA and BayesB the parameters of the Student's t prior density were defined to produce the desired genetic variance (Meuwissen *et al.* 2001). The Xu (2003) method is otherwise similar to BayesA, except that the prior parameters are estimated instead of setting into constant values. Similar modifications of BayesB have been considered by *e.g.* Yi and Xu (2008) and Habier *et al.* (2011). Under the Bayesian LASSO the prior parameters are more commonly estimated from data (*e.g.* Yi and Xu 2008; de los Campos *et al.* 2009; Sun *et al.* 2010; Shepherd *et al.* 2010) than given as constants (Xu 2010).

While the Bayesian models have proven workable, efficient and flexible, the tremendous number of markers in the modern genome-wide data sets make the computational methods traditionally connected to Bayesian estimation, *e.g.* Markov Chain Monte Carlo (MCMC), quite slow and cumbersome. For the same models fast alternative estimation procedures have been proposed, most commonly based on estimation of the maximum point of the posterior density (MAP-estimate), rather than the whole posterior distribution, by expectation-maximization (EM) algorithm (Dempster *et al.* 1977; McLachlan and Krishnan 1997; for the methods see *e.g.* Yi and Banerjee 2009; Hayashi and Iwata 2010; Figueiredo 2003; Sun *et al.* 2010; Xu 2010; Meuwissen *et al.* 2009; Shepherd *et al.* 2010; Lee *et al.* 2010).

2 Objectives of the study

The objectives of this work are to 1) better understand the behavior of the different Bayesian multilocus association models, especially under the *maximum a posteriori* estimation context, and to obtain further information on the instances in which different methods work best, 2) seek connections between the different Bayesian models and try to see the different model variants as special cases or sub-models of a common model framework, 3) pay special attention to the significance of the parametrization and hierarchical structure of the model for elegant derivation and convergence properties of the estimation algorithm, and 4) to develop a flexible and versatile Bayesian multilocus association model framework, along with an efficient parameter estimation machinery, that can be utilized in phenotype prediction, genomic breeding value estimation and QTL (quantitative trait loci) detection and effect estimation from a variety of genome-wide data.

The original papers I–III contribute to the objectives in the following manner.

In I we lay the foundation for our Bayesian model framework, examine the behavior and predictive performance of different sub-models and prior densities, including G-BLUP, and present a generalized expectation-maximization algorithm (GEM) for the parameter estimation.

In II we apply selected parts of the model framework in QTL mapping context and, in particular, consider the impact of an additional polygenic component for the performance of the model and the GEM-algorithm.

In III we generalize the model framework and the GEM-algorithm for ordered categorical and censored Gaussian phenotypes.

3 Hierarchical Bayesian model

In Bayesian inference the learning from data is based on updating the prior belief concerning the model parameters into the posterior belief by applying the Bayes' theorem. Let $p(\Theta)$ denote the joint prior density for the unknown parameters and $p(data|\Theta)$ the likelihood of the data given those parameters. Now the posterior density for the unknown parameters, given the data, is acquired from the Bayes' formula

$$p(\Theta|data) = \frac{p(data|\Theta)p(\Theta)}{p(data)} \propto p(data|\Theta)p(\Theta),$$

where the normalizing constant $p(data) = \int_\Theta p(data|\Theta)p(\Theta)d\Theta$ is the marginal likelihood of the data. As the marginal likelihood has a constant value, it is usually omitted from the computation, and the joint posterior density is considered to be proportional to the product of the likelihood and the joint prior density. In addition to the prior conception of the parameter values, the joint prior density expresses the mutual relationships of the parameters, *e.g.* whether the parameters are considered *a priori* independent or conditional to some other parameters. This definition is denoted as the hierarchical structure of the Bayesian model. Let *e.g.* the parameter vector be $\Theta = (\theta_1, \theta_2)$, and let θ_1 be *a priori* dependent on θ_2. Now the joint prior is given by $p(\Theta) = p(\theta_1|\theta_2)p(\theta_2)$, and the dependent parameter θ_1 is said to be located on a lower layer of the model hierarchy.

In its complete form our hierarchical Bayesian model framework, depicted as a directed acyclic graph in Figure 3.1, consists of two separate parts, the linear Gaussian model and the threshold model. Under the linear Gaussian model the phenotype measurements are assumed to be continuous and follow a Gaussian density, while the additional threshold model handles binary, ordinal and censored Gaussian observations. The hierarchical model has a total of six layers, two of which are optional. The observed data, located on the 1st and 2nd layers in the graph, comprises phenotype and genotype information and, optionally, a known pedigree of a sample of related individuals. The continuous Gaussian phenotypes, denoted by a vector $\mathbf{y}$, and the genetic data matrix $\mathbf{X}$ consisting the genotypes of biallelic SNP markers, are located on the "observed data" layer of the linear Gaussian model. As the binary, ordinal and censored Gaussian observations are handled via a latent variable parametrization, they are located on the "optional observed" layer of the threshold model in Figure 3.1. The possible pedigree information is given in a form of an additive genetic relationship matrix (Lange 1997), located on the "optional observed" layer

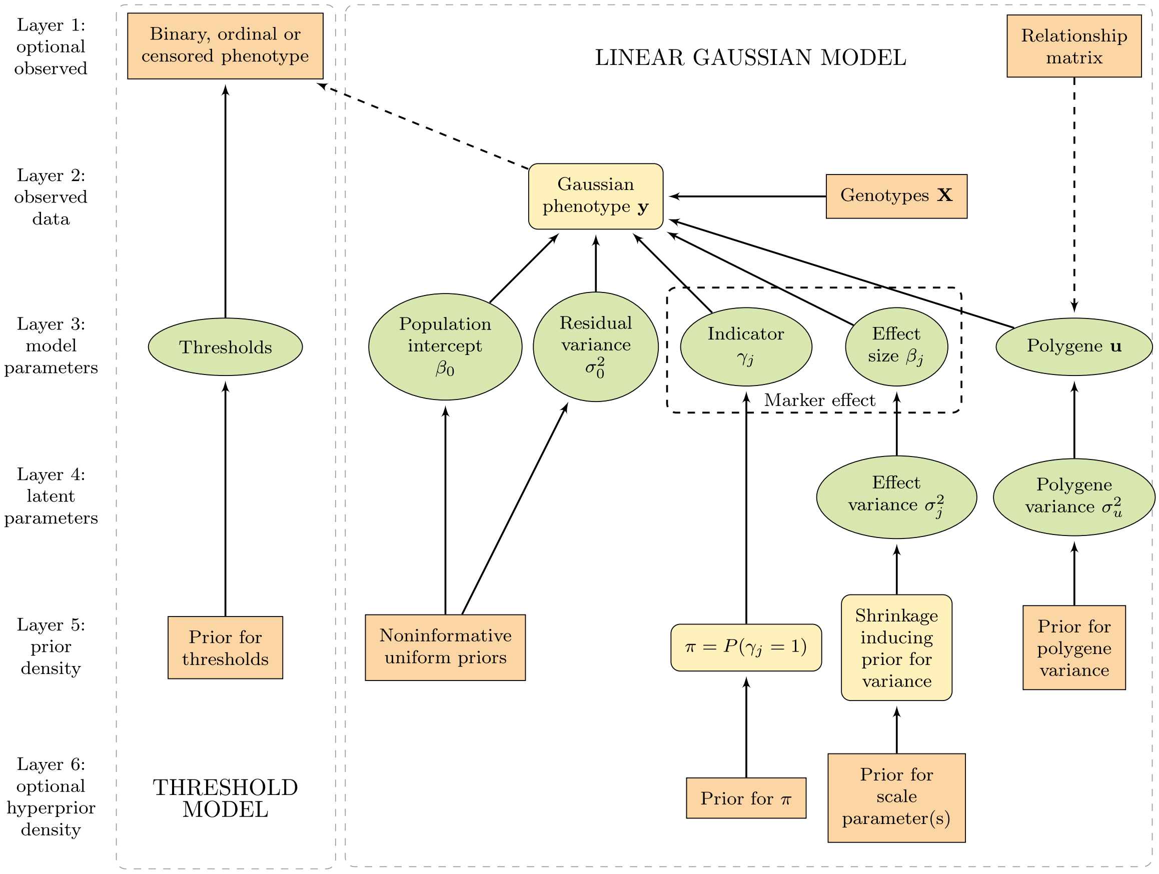
Layer 1: optional observed
Layer 2: observed data
Layer 3: model parameters
Layer 4: latent parameters
Layer 5: prior density
Layer 6: optional hyperprior density
Binary, ordinal or censored phenotype
Thresholds
Prior for thresholds
THRESHOLD MODEL
LINEAR GAUSSIAN MODEL
Relationship matrix
Gaussian phenotype $\mathbf{y}$
Genotypes $\mathbf{X}$
Population intercept β_0
Residual variance σ_0^2
Indicator γ_j
Effect size β_j
Marker effect
Polygene $\mathbf{u}$
Effect variance σ_j^2
Polygene variance σ_u^2
Noninformative uniform priors
$\pi = P(\gamma_j = 1)$
Shrinkage inducing prior for variance
Prior for polygene variance
Prior for π
Prior for scale parameter(s)

Figure 3.1: Hierarchical structure of the model framework. The ellipses represent random parameters and rectangles fixed values, while the round-cornered rectangles may be either, depending on the selected model. Solid arrows indicate statistical dependency and dashed arrows functional relationship. The background boxes indicate the main modules of the model framework.

in the directed acyclic graph (Figure 3.1) to represent its non-compulsory nature.

In the following sections we first will consider the linear Gaussian model part, and only after that focus on the threshold model for the discrete or censored data.

3.1 Gaussian likelihood

In the center of a Bayesian model there is the likelihood function of the data given the model parameters. The likelihood is based on the probability model (sometimes called the sampling model) determining how the independent variables or traits are connected to the explanatory variables. In our model framework the Gaussian phenotypes are connected to the marker and pedigree information with a linear Gaussian association model (see Figure 3.1)

$$\mathbf{y} = \beta_0 + \mathbf{X}\mathbf{\Gamma}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{y}$ denotes the phenotypic records of n individuals, β_0 is the population intercept, and $\boldsymbol{\varepsilon}$ corresponds to the residuals, assumed normal and independent, $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}_n\sigma_0^2)$. If necessary, the intercept β_0 can be easily replaced with a vector of environmental variables. The second term on the right hand side of the equation (3.1) comprises the observed genotypes $\mathbf{X}$ and the allele substitution effects $\mathbf{\Gamma}\boldsymbol{\beta}$. The observed genotypes of the p biallelic SNP markers are coded with respect to the number of the rare alleles (0,1 and 2) and standardized to have null mean and unity variance. In the complete model the allele substitution effect (see "Marker effect" in Figure 3.1) is modeled following Kuo and Mallick (1998) as a product of the size of the effect and a variable indicating whether the marker is linked to the phenotype. In the equation (3.1), $\boldsymbol{\beta}$ denotes the additive effects sizes, and $\mathbf{\Gamma}$ is a diagonal matrix of indicator variables, whose jth diagonal element γ_j has value 1 if the jth SNP marker is included in the model, and 0 otherwise. As depicted in Figure 3.1, the indicator and the effect size are considered *a priori* independent. The term $\mathbf{u}$ in the equation (3.1) denotes the additive polygenic effects due to the combined effect of infinite number of loci, and $\mathbf{Z}$ is a design matrix connecting the polygenic effects to the observed phenotypes.

The individuals, or their phenotypic values y_i, are assumed conditionally independent given the genotype information $\mathbf{X}$ and the polygenic effect $\mathbf{u}$. This assumption and the described linear marker association model (3.1) gives a multivariate normal likelihood

$$\begin{aligned} p\left(\mathbf{y}\,|\,\beta_0, \sigma_0^2, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \mathbf{u}, \mathbf{X}, \mathbf{Z}\right) &\propto \det(\mathbf{I}_n\sigma_0^2)^{-1/2} \\ \times \exp\left(-\frac{1}{2}(\mathbf{y}-\beta_0-\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'(\mathbf{I}_n\sigma_0^2)^{-1}(\mathbf{y}-\beta_0-\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\right) & \qquad (3.2) \end{aligned}$$

for the phenotypes given the parameter vector. Due to the independence of the observations, the likelihood can be interpreted also as an univariate normal $\mathrm{N}(\beta_0 + \sum_{j=1}^{p}\gamma_j\beta_j x_{ij} + u_i, \sigma_0^2)$ given a single observation y_i and the appropriate parameters. The parameters of the multilocus association model that are present in the likelihood function are located in the "model parameters" -layer of the linear Gaussian model in Figure 3.1.

3.2 Shrinkage inducing priors

The second essential component of a Bayesian model consists of the prior densities for the model parameters. The prior for a given parameter represents the *a priori* understanding of the plausibility of different parameter values. In some cases there is no reason to believe that one parameter value would be more plausible than another, which conception is expressed with a flat or an uninformative prior density, *e.g.* by setting $p(\beta_0) \propto 1$ and $p(\sigma_0^2) \propto 1/\sigma_0^2$ (note the "Noninformative uniform priors" at layer 5 in Figure 3.1). In some cases, however, the prior density plays a most important role in the model operation.

A central feature of handling an oversaturated model is the selection or regularization of the excess predictors. In the Bayesian context the regularization is included into the model by specifying such a prior density for the regression coefficients, that it represents the *a priori* understanding that the majority of the predictors have only a negligible effect, while there are a few predictors with possibly large effect sizes. A prior that would evince this idea should consist of a probability mass centered near zero and a probability mass distributed over the nonzero values, including a reasonably high probability for large values. The probability density functions we have used for imitating this "spike and slab" shape are Student's t (following *e.g.* Meuwissen *et al.* 2001; Xu 2003) and Laplace densities (following *e.g.* Park and Casella 2008; de los Campos *et al.* 2009), either alone or combined with a point mass at zero (*e.g.* Meuwissen *et al.* 2001; Shepherd *et al.* 2010).

In our full model framework, (3.1) and Figure 3.1, the mixture prior with the point mass at zero is accomplished by adding a dummy variable to indicate whether the effect of a given predictor variable is included into the model or not. Following Kuo and Mallick (1998) the marker effects are modeled as a product of the indicator variable γ_j and the effect size β_j, which are considered *a priori* independent, hence the joint prior of the marker effect becomes simply $p(\gamma_j \beta_j) = p(\gamma_j)p(\beta_j)$, where $p(\gamma_j)$ is a Bernoulli density with a prior probability $\pi = \mathrm{P}(\gamma_j = 1)$ for a marker to be linked to the trait and $p(\beta_j)$ is the prior density for the effect size.

3.2.1 Hierarchical formulation of the prior densities

The Student's t and the Laplace distribution can both be expressed as a scale mixture of normal distributions with a common mean and effect specific variances. The hierarchical formulation of a Student's t-distribution with ν degrees of freedom, location μ and scale τ^2 is a scale mixture of normal densities with mean μ and variances following a scaled inverse-χ^2 distribution with ν degrees of freedom and scale τ^2,

$$\left.\begin{array}{rcl} \beta_j|\sigma_j^2 & \sim & \mathrm{N}(\mu, \sigma_j^2) \\ \sigma_j^2|\nu, \tau^2 & \sim & \text{Inv-}\chi^2(\nu, \tau^2) \end{array}\right\} \Longrightarrow \beta_j \sim t_\nu(\mu, \tau^2),$$

while a Laplace density with location μ and rate λ can be presented in a similar manner, the mixing distribution now being an exponential one,

$$\left.\begin{array}{rcl} \beta_j|\sigma_j^2 & \sim & \mathrm{N}(\mu, \sigma_j^2) \\ \sigma_j^2|\lambda^2 & \sim & \mathrm{Exp}(\lambda^2/2) \end{array}\right\} \Longrightarrow \beta_j \sim \mathrm{Laplace}(\mu, \lambda).$$

The hierarchical representation of the prior densities bears a twofold advantage (I). First, the derivation of the fully conditional posterior densities, and hence the derivation of the estimation algorithm, simplifies greatly. Within MCMC world, the hierarchical formulation of the prior densities, also known as model- or parameter expansion, is a well known device to simplify computations by transforming the prior into a conjugate and thus enabling Gibbs sampling. Conjugacy of a prior distribution means that the fully conditional posterior probability distribution of a given parameter will be of same type as the prior distribution of that parameter, and hence we are guaranteed to get a closed form fully conditional posterior with a known probability density function. The hierarchical formulation of a prior density is also known to accelerate convergence of a MCMC sampler by adding more working parts and therefore more space for the random walk to move (see *e.g.* Gilks *et al.* 1996; Gelman *et al.* 2004; Gelman 2004). In *maximum a posteriori* (MAP) estimation, on the other hand, a commonly adopted approach to try and simplify the model is to integrate out the effect variances. However, the conjugacy maintained by preserving the intermediate variance layer (layer 4 in Figure 3.1) is a valuable feature also for MAP-estimation, as it enables the straightforward derivation of the fully conditional posterior density functions. Expressed as a scale mixture, the Student's t distribution leads to conjugate priors for normal likelihood parameters, and hence is a perfect choice for a conjugate analysis. Although the decomposition of the Laplace prior does not provide conjugacy, it leads to a tractable fully conditional posterior density for the inverse of the effect variance.

Second, the estimation algorithm is likely to behave better under a hierarchical model. Even though the marginal distributions of the marker effects are mathematically equivalent in hierarchical and non-hierarchical models, we noted in I that the parametrization and model structure alter the properties and behavior of the model, and thus have influence on the mixing and convergence properties of an estimation algorithm, and also on the values of the actual estimates. We noted in I that in some cases the hierarchical Laplace model was clearly more accurate than its non-hierarchical counterpart. Also, contrary to the non-hierarchical version, the hierarchical Laplace model worked without the additional indicator variable, *i.e.* without a zero-point-mass in the prior of the marker effects. This simplification of the model leads not only to more straightforward implementation and faster estimation, but also to easier and more accurate selection of prior parameters.

3.3 Sub-models

As mentioned above, we like to consider the full model in Figure 3.1 as a framework incorporating a set of model variants, or sub-models, embodying different components of the model framework. In I we covered a multitude of such variants, and also showed how the model variants correspond to the Bayesian phenotype prediction and genomic breeding value estimation methods proposed in the literature.

The non-compulsory components of the multilocus association model comprise the polygenic component, the indicator variable and the 6th, ”optional hyperprior” layer. The selection between the Student's t and the Laplace prior densities forms one means of modifying the prior density assigned for the marker effects, while the inclusion/exclusion of the indicator and the hyperprior layer forms another. The polygenic component, on the other hand, is clearly an external addition to the multilocus association model.

3.3.1 Polygenic component

The polygenic component $\mathbf{u}$ is included into the model to represent the genetic variation possibly not captured by the SNP markers and to take account for putative residual dependencies between individuals (Yu *et al.* 2006). The sample or population structure is included into the model as the covariance matrix of the multivariate normal prior density given for the polygenic effect $\mathbf{u}|(\sigma_u^2, \mathbf{A}) \sim \mathrm{MVN}(\mathbf{0}, \sigma_u^2\mathbf{A})$, where σ_u^2 is the polygenic variance component and $\mathbf{A}$ is the genetic relationship matrix. The genetic relationship matrix is either a pedigree based additive genetic relationship matrix (see Lange 1997) (I and II), or, if there is no pedigree available, a finite locus approximation based on the markers not included in the actual multilocus association model (a genomic relationship matrix) (II). The polygenic variance component σ_u^2 has been given an Inverse-$\chi^2(\nu_u, \tau_u^2)$ prior distribution with suitable data specific parameter values.

On the basis of the existing literature the need for an additional polygenic component within a multilocus association model is unclear. Many authors have found the polygenic component irrelevant (*e.g.* Calus and Veerkamp 2007; Pikkuhookana and Sillanpää 2009), while *e.g.* de los Campos *et al.* (2009) and Lund *et al.* (2009) see it as a necessary. In I and II we examined the importance of the additional polygenic component in genomic selection and in association mapping context, respectively, with both simulated and real data. Within these works the estimates of the polygenic component were negligible, and had no influence neither in the prediction accuracy (I) nor in the gene location ability (II) of the model. None of the Bayesian multilocus models seemed to benefit from addition of the polygenic component with neither simulated (I and II) nor real data (I), the phenotype of the latter most likely being quite polygenic in nature. The polygenic component did not find extra information even when the task was made as easy as possible by generating the polygenic component of the data by using the same relationship matrix which was also used in the analyses (II). Therefore, to our experience, the polygenic component can safely be omitted from the multilocus association model (3.1).

3.3.2 Indicator

The indicator variable is added to the model framework to participate as a source of extra shrinkage in a mixture prior alongside the Student's t or the Laplace density. The usefulness of the indicator variable depends on the other source of shrinkage in the model. As mentioned above, the hierarchical Laplace model does not seem to require the additional point mass at zero, on the contrary the model efficiency sustains damage if the indicator is added (Tables 2–5 in I). On the other hand, the Student's t model clearly benefits from the additional point mass. The latter observation is in strict concordance with the existing literature, as the superiority of BayesB (Student's t plus indicator) (Meuwissen *et al.* 2001) over BayesA (only Student's t) can be considered as common knowledge.

While the main purpose of the indicator variable within our model framework is to participate in the mixture prior with the Student's t or Laplace densities, in II we have considered a pure indicator model. Under the Indicator model proposed in II, the prior for the effect sizes β_j is Gaussian with zero mean and a predetermined variance, and therefore the prior for the marker effects $\gamma_j \beta_j$ is a mixture of a Gaussian density and a point mass at zero. As the Gaussian prior introduces a constant shrinkage to the estimates, and hence the variable selection relies solely on the indicator, a Bayes factor based on the values of the indicators can be used in determining the significance of a marker effect. Contrary to phenotype or breeding value prediction, in gene mapping the significance of the individual marker effects is of importance. Nevertheless, the Indicator model in II is mainly considered as a curiosity and a proof of the power of a multilocus association treatment, as even an extremely simple multilocus association method may exceed the performance of a most sophisticated single marker method (Figure 1, A and B in II).

The indicator has a Bernoulli prior with a prior probability $\pi = \mathrm{P}(\gamma_j = 1)$ for the SNP j contributing to the trait. The value given for the probability π also represents our prior assumption of the proportion of the SNP markers that are linked to the trait. However, as the indicator affects the shrinkage of the marker effects concurrent with the shrinkage generated by the Student's t or the Laplace density, the parameters assigned for these densities affect the selection of π.

3.3.3 Hyperprior

The optional hyperprior layer (the 6th layer in Figure 3.1) composes another facultative part of the model framework. The parameters of the prior densities (layer 5 in Figure 3.1) can be either predetermined or estimated simultaneously to the model parameters. As the prior densities for the effect size and the indicator are responsible for the regularization of the excess variables in the model, the impact of the parameter values of these priors is greater than of the other prior densities in the model. Therefore the putative estimation of the prior parameters is limited to these two parameters. The estimation of the prior parameters is depicted in Figure 3.1 by considering the priors for the indicator and the effect variance as random variables, and adding the 6th layer into the model. If the parameters for the prior densities are considered fixed, the optional hyperprior layer is absent from the model. The fixed prior parameter values can be determined *e.g.* by cross validation or by Bayesian information criterion (see Sun *et al.* 2010). It is noteworthy, that even if the prior parameters are estimated from the data, *i.e.* the 6th layer is present in the model, the need for predetermined values does not vanish, but simply passes to the next layer of the model hierarchy. Hence, inevitably, at the very bottom of the model hierarchy the user has to determine some values prior to the actual parameter estimation.

The hyperprior given for the effect size is a conjugate Gamma(κ, ξ) density for the scale τ^2 of the inverse-χ^2 density under the Student's t model, or, respectively, for the rate λ^2 of the exponential density under the Laplace model. There is neither conjugate prior nor closed form posterior density available for the degrees of freedom parameter of the Student's t model, and hence we have decided to consider it as fixed (I). For the indicator variable, the prior probability $\pi = \mathrm{P}(\gamma_j = 1)$ of the marker j to be linked to the trait, is estimated with either an uninformative uniform Beta(1,1), or an informative Beta(a, b) density. The informative beta prior embodies our *a priori* assumed belief of the proportion of significant markers by considering a as the number of markers assumed to be linked to trait and b as the number of markers not to be linked (*i.e.* $b = p - a$, p being the number of markers in the data set).

3.3.4 Student's *t vs.* Laplace prior

In the original work I one of our main interests was to consider the pros and cons of the Student's t and Laplace prior densities. The advantage of the hierarchically formulated Student's t density as a prior is the extremely easy derivation of the fully conditional posterior densities. Although the hierarchical Laplace prior also leads to tractable fully conditional posterior functions, the derivation of the posterior for the effect variance is clearly more complicated than with the Student's t density. However, the Student's t prior has some shortcomings too. The first problem we encountered with the Student's t model was the estimation of the parameters for the prior densities (5th layer in the Figure 3.1). We tried numerous hyperpriors for the effect variance and the indicator, but it appeared to be impossible to select ones leading to a reasonable estimate. Hence, after several attempts, we decided on treating the prior parameters of the Student's t model as given. Under the Laplace model there was no such complications, and the prior parameters of the Laplace model are estimated from the data. Therefore, in the Laplace model the 6th layer of Figure 3.1 is always included in the model, while in the Student's t model it is always excluded from the model.

Due to its shape, the shrinking ability of the Student's t prior is weaker than of the Laplace prior. While the hierarchical Laplace prior worked fine without the additional indicator variable, the Student's t prior required the additional point mass at zero in order to provide a strong enough shrinkage (Tables 2–5 in I). As pointed out previously, a low number of parameters is a desirable characteristic in a model. Apart from a single data set (table 2 in I), the prediction accuracy of the Laplace model was higher compared to the Student's t model (Tables 3-5 in I). The better performance of the Laplace model may be partially due to the easier and hence more accurate prior selection, partially due to the more favorable shape of the density itself. Also, as the prior parameters for the effect variance can be estimated, and hence there is an additional layer in the hierarchical model, the model may be more robust to the given hyperprior parameter values. Altogether, on the basis of our findings in I, we feel that the hierarchical Laplace model appears to have an advantage over the Student's t model, and therefore decided to concentrate on the former in II and in III.

3.3.5 Bayesian LASSO and its extensions

The hierarchical Bayesian model with a Laplace prior density is commonly denoted as the Bayesian LASSO (Park and Casella 2008) since it leads to a nearly identical estimate as the frequentist LASSO by Tibshirani (1996). The Bayesian LASSO has been further modified by several authors, including Yi and Xu (2008), Mutshinda and Sillanpää (2010), Sun *et al.* (2010) and Fang *et al.* (2012).

In II we considered a modification of the Bayesian LASSO introduced by Mutshinda and Sillanpää (2010) called the Extended Bayesian LASSO (EBL). Following common hierarchical Bayesian LASSO, the Laplace prior is expressed as a scale mixture of normal densities with exponential mixing distribution, so the EBL assigns a normal prior with independent locus-specific variances to the regression parameters given the locus variances $\beta_j|\sigma_j^2 \sim \mathrm{N}(0, \sigma_j^2)$, and further an exponential prior to the variances $\sigma_j^2|\lambda_j^2 \sim \mathrm{Exp}(\lambda_j^2/2)$. Unlike Bayesian LASSO, the regularization parameters λ_j^2 of

EBL are locus specific, and can be decomposed by setting $\lambda_j = \delta\eta_j$, where δ represents the model sparseness common to all loci, and η_j is a locus-specific deviation representing the shrinkage working at locus j. Now the common Bayesian LASSO can be seen as a special case of EBL with the locus specific component set to $\eta_j = 1 \; \forall \; j$. Setting the common shrinkage parameter $\delta = 1$ would lead to the Improved Bayesian LASSO proposed by Fang *et al.* (2012).

3.3.6 Bayesian G-BLUP

In addition to the multilocus association model, in I and III we have considered a Bayesian version of the genomic- or G-BLUP, a classical BLUP model where the numerator relationship matrix, estimated from the pedigree, is replaced by a genomic marker-based relationship matrix

$$\mathbf{y} = \beta_0 + \mathbf{Zu} + \boldsymbol{\varepsilon}. \tag{3.3}$$

In the model framework in Figure 3.1 the G-BLUP can be seen as a mirror image of the multilocus association model without the polygenic component, as here we have the polygene without the marker effects. The likelihood of the data under the G-BLUP is simply a multivariate normal with mean $\beta_0 + \mathbf{Zu}$ and covariance $\mathbf{I}_n\sigma_0^2$. Prior for the genetic values $\mathbf{u}$ and the population intercept β_0 are conjugate multivariate normal $\text{MVN}(\mathbf{0}, \mathbf{G}\sigma_u^2)$ and uniform, respectively, $\mathbf{G}$ being the genomic relationship matrix. The variances σ_0^2 and σ_u^2 have inverse-χ^2 priors, uninformative $p(\sigma_0^2) \propto 1/\sigma_0^2$ and a level $\text{Inv-}\chi^2(\nu_u, \tau_u^2)$, respectively.

Under the G-BLUP the genetic marker data is incorporated into the model in a form of a genomic relationship matrix. There are numerous methods of generating the genomic relationship matrix, we have used the second method described in VanRaden (2008). This method is based on the identity by state (IBS) of the marker genotypes, and hence it measures the realized relationship between the individuals.

The Bayesian approach differs from the frequentist G-BLUP in terms of handling the variance components. While the frequentist methods commonly estimate the genomic breeding values with known variance components, in a Bayesian approach the variance components are estimated simultaneously to the breeding values (Hallander *et al.* 2010). Therefore the Bayesian inference is always based on variances that are up-to-date and specific to the analyzed trait, letting also the uncertainty of the variance components to be incorporated into the estimates of the breeding values. Even though *e.g.* ASREML (Gilmour *et al.* 2009) estimates the variance components from the data, and hence satisfies the up-to-date criterion, the variances are not estimated simultaneously to the breeding values, instead, the pre-estimated variance components are considered constant while estimating the breeding values.

3.4 Fully conditional posterior densities

As depicted in the Figure 1, the model parameters $\beta_0, \sigma_0^2, \boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\mathbf{u}$, located at the 3rd layer, are considered *a priori* independent of each other. The prior independence of the indicator and the effect size, as suggested by Kuo and Mallick (1998), leads to the most straightforward parametrization of a mixture prior for the effects. In conjunction with the conjugate, or otherwise well chosen, prior densities it enables an easy derivation of a closed form fully conditional posterior distribution for every parameter of the model framework.

The joint posterior distribution of the parameters, given the data, is proportional to the product of the joint prior and the likelihood. We can easily extract the fully conditional posterior densities of individual parameters from the joint posterior by handling all other parameters as constants and leaving them out, and hence selecting only the terms including the parameter in question. For example, the fully conditional posterior distribution of a single regression coefficient β_j, given all other parameters and the data, is derived from the joint distribution simply by selecting only the terms including β_j, *i.e.* the likelihood and the conditional prior $p(\beta_j|\sigma_j^2)$.

Under the full multilocus association model (3.1) we get the following, closed form, fully conditional posterior distributions for the model parameters (for simplicity: $\star =$ "the data, and the parameters except the one in question"):

$$\beta_0 \,|\, \star \sim \mathrm{N}\Big(\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\gamma_j\beta_j x_{ij} - u_i),\ \frac{\sigma_0^2}{n}\Big), \tag{3.4}$$

$$\sigma_0^2 \,|\, \star \sim \text{Inv-}\chi^2\Big(n,\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\gamma_j\beta_j x_{ij} - u_i)^2\Big), \tag{3.5}$$

$$\begin{aligned} \beta_j \,|\, \star &\sim \mathrm{N}(\mu_j, s_j^2), \text{ where} \\ \mu_j &= \sum_{i=1}^{n}\gamma_j x_{ij}\Big(y_i - \beta_0 - \sum_{l\neq j}\gamma_l\beta_l x_{il} - u_i\Big) \Big/ \Big(\sum_{i=1}^{n}(\gamma_j x_{ij})^2 + \frac{\sigma_0^2}{\sigma_j^2}\Big), \\ s_j^2 &= \sigma_0^2 \Big/ \Big(\sum_{i=1}^{n}(\gamma_j x_{ij})^2 + \frac{\sigma_0^2}{\sigma_j^2}\Big), \end{aligned} \tag{3.6}$$

$$\begin{aligned}
\mathbf{u}\,|\,\star \quad &\sim \quad \text{MVN}(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), \text{ with} && (3.7)\\
\boldsymbol{\mu}_u \quad &= \Big(\mathbf{Z}'\mathbf{Z} + \frac{\sigma_0^2}{\sigma_u^2}\mathbf{A}^{-1}\Big)^{-1}\mathbf{Z}'\Big(\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\Big), \text{ and}\\
\boldsymbol{\Sigma}_u \quad &= \Big(\frac{1}{\sigma_0^2}\mathbf{Z}'\mathbf{Z} + \frac{1}{\sigma_u^2}\mathbf{A}^{-1}\Big)^{-1},
\end{aligned}$$

$$\begin{aligned}
\gamma_j\,|\,\star \quad &\sim \quad \text{Bernoulli}\Big(\frac{\pi\mathcal{R}_j}{(1-\pi)+\pi\mathcal{R}_j}\Big) \text{ , where} && (3.8)\\
\mathcal{R}_j \quad &= \frac{p(\mathbf{y}|\gamma_j = 1, \boldsymbol{\theta}_{-\gamma_j})}{p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\theta}_{-\gamma_j})}\\
&= \exp\Big(\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\Big(2\beta_j x_{ij}\Big(y_i - \beta_0 - \sum_{l\neq j}\gamma_l\beta_l x_{il} - u_i\Big) - \Big(\beta_j x_{ij}\Big)^2\Big)\Big).
\end{aligned}$$

The corresponding fully conditional posterior densities for the different submodels can be derived from the above by eliminating the obsolete model components. If the polygenic component $\mathbf{u}$ is not included in the submodel, we set $u_i = 0$ for all i in the other posteriors. Correspondingly, if the indicator $\boldsymbol{\gamma}$ is absent, $\gamma_j = 1$ for all j. In the Bayesian G-BLUP (3.3) the marker effect in its entirety is absent since there is no genotype matrix ($\mathbf{X}$) present in the model, and the numerator relationship matrix $\mathbf{A}$ is replaced by the genomic relationship matrix $\mathbf{G}$. The residual variance σ_0^2 is updated only if the Gaussian phenotype is fully observed, otherwise (when the threshold module is present in the model) it is set to unity (this is discussed at the following section).

The fully conditional posteriors for the latent variance parameters (layer 4 in Figure 3.1) are as follows. Under the Student's t model the fully conditional posterior for the effect variance is

$$\sigma_j^2\,|\,\star \quad \sim \quad \text{Inv-}\chi^2\Big(\nu+1, \frac{\beta_j^2 + \nu\tau^2}{\nu+1}\Big), \tag{3.9}$$

and under the Laplace model the fully conditional posterior for the inverse of the effect variance is an inverse-Gaussian (Chhikara and Folks 1989)

$$\frac{1}{\sigma_j^2}\,|\,\star \quad \sim \quad \text{Inverse-Gaussian}\Big(\frac{\lambda}{|\beta_j|}\text{ , }\lambda^2\Big), \tag{3.10}$$

the parametrization of an Inverse-Gaussian(μ', λ') density with mean μ' and shape λ' being

$$f(x|\mu', \lambda') \quad \propto \quad x^{-3/2}\exp\left(-\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x}\right).$$

The fully conditional posterior density for the variance of the polygene is

$$\sigma_u^2 \,|\, \star \quad \sim \quad \text{Inv-}\chi^2\Big(\nu_u + N,\; \frac{\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u \tau_u^2}{\nu_u + N}\Big), \tag{3.11}$$

where N denotes the total number of individuals in the learning and prediction sets, and the numerator relationship matrix $\mathbf{A}$ is replaced by the genomic relationship matrix $\mathbf{G}$, when necessary.

When the hyperprior layer is included in the model, the prior parameters (layer 5 in Figure 3.1) have the following fully conditional posterior densities. The fully conditional posterior density for the probability $\pi = \mathrm{P}(\gamma_j = 1|\star)$ of a given marker to be linked to the trait is given by

$$\pi \,|\, \star \quad \sim \quad \text{Beta}\Big(a + \sum_{j=1}^{p} \gamma_j \,,\, b + p - \sum_{j=1}^{p} \gamma_j\Big). \tag{3.12}$$

Under the Student's t model, the fully conditional posterior density for the scale of the inverse-χ^2 distribution is

$$\tau^2 \,|\, \star \quad \sim \quad \text{Gamma}\Big(\frac{(\kappa - 1)p\,\nu}{2} + 1 \,,\, \frac{\nu + 1}{2\,\xi} \sum_{j=1}^{p} \frac{1}{\sigma_j^2}\Big), \tag{3.13}$$

however, as mentioned above, in practice we have not included the optional hyperprior layer into the Student's t model. Under the Laplace model, or the Bayesian LASSO, the fully conditional posterior density for the regularization parameter is given by

$$\lambda^2 \,|\, \star \quad \sim \quad \text{Gamma}\Big(\kappa + p \,,\, \xi + \sum_{j=1}^{p} \frac{\sigma_j^2}{2}\Big). \tag{3.14}$$

3.5 Threshold model

In this section we shall consider the threshold model part of the full model framework depicted in Figure 3.1. We assume that the observed phenotype $\mathbf{w}$ consists of either binary, ordered categorical or censored Gaussian observations, and that the ordered categorical variable has arisen as an underlying normally distributed continuous response $\mathbf{y}$ is rendered discrete with known number of thresholds at unknown positions. Now the underlying Gaussian response $\mathbf{y}$ can be explained by the genetic factors with the multilocus association model (3.1) or one of the sub-models, including the G-BLUP (3.3). As the Gaussian response $\mathbf{y}$ is unobservable, in order to avoid overparametrization the residual variance component σ_0^2 of the linear model is set to unity.

Given the value of the continuous, normally distributed latent variable y_i, the binary or ordinal response w_i has value $k \in \{1, \ldots, K\}$ with a probability

$$\mathrm{P}(w_i = k \mid y_i, t_{k-1}, t_k) = \begin{cases} 1, & \text{when } \ t_{k-1} < y_i < t_k \\ 0, & \text{otherwise,} \end{cases} \tag{3.15}$$

where t_{k-1} and t_k are the thresholds delimiting the kth category. If the ordinal variable has K categories, there will be $K+1$ thresholds, such that $\mathbf{t} = \{(t_0, t_1, \ldots, t_K) | t_0 < t_1 < \cdots < t_K, t_0 = -\infty, t_1 = 0, t_K = \infty\}$. The $K-2$ of the thresholds $\mathbf{t}^\star = \{(t_2, \ldots, t_{K-1}) | t_2 < \cdots < t_{K-1}\}$ are considered unknown, and are estimated simultaneously to the model parameters. With a binary response ($K = 2$) there obviously are no unknown threshold values.

As defined in (3.15), conditionally on the underlying response and the thresholds, the observed ordinal phenotype w_i is known with certainty and hence the likelihood is degenerated into a constant value, zero or one. Under the threshold model, the prior density for the latent variable y_i corresponds to the likelihood of the Gaussian response under the linear Gaussian model (3.1) or (3.3) with residual $\varepsilon \sim \mathrm{N}(0, 1)$. Due to the degenerate likelihood of the observed phenotype w_i, the fully conditional posterior density of the latent Gaussian variable y_i, given the value of the observed phenotype, corresponds the prior density of y_i when $t_{k-1} < y_i < t_k$ and is zero otherwise. Hence, the fully conditional posterior density of y_i is a truncated normal distribution (truncated at points t_{k-1} and t_k) with a density function (for simplicity, the $\star$ denotes the data and all other parameters)

$$p(y_i | \star) = \frac{\phi(y_i - \mathbb{E}(y_i))}{\Phi(t_k - \mathbb{E}(y_i)) - \Phi(t_{k-1} - \mathbb{E}(y_i))}, \tag{3.16}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions, respectively, while $\mathbb{E}(y_i)$ is the linear predictor of the model (3.1) or (3.3).

Following Sorensen *et al.* (1995) the prior for the $K-2$ unknown thresholds $\mathbf{t}^\star = (t_2, \ldots, t_{K-1})$ has been given as order statistics from an Uniform$(0, t_{max})$ distribution,

$$p(\mathbf{t}^\star) = (K-2)! \Big(\frac{1}{t_{max}}\Big)^{K-2} \text{ for } \ 0 < t_2 < \ldots t_{K-1} < t_{max}, \text{ and 0 otherwise.} \tag{3.17}$$

Note, that the threshold values $\mathbf{t}^\star$ appear in the prior density only at the definition of the support of the distribution. As the terms not including the parameter are discarded as constants from the fully conditional posterior, the support definition is all that passes from the prior to the posterior.

Therefore, the fully conditional posterior density for a t_k is given by the likelihood of the observed ordinal phenotype $\mathbf{w}$, within the set of values determined by the prior density of $\mathbf{t}^\star$,

$$p(t_k|\star) \propto \prod_{i=1}^{n} \mathrm{P}(w_i = k)^{\mathrm{I}(w_i=k)} \mathrm{P}(w_i = k+1)^{\mathrm{I}(w_i=k+1)}$$

$$\propto \prod_{i=1}^{n} \mathrm{P}(t_{k-1} < y_i < t_k | t_{k-1}, t_k)^{\mathrm{I}(w_i=k)} \mathrm{P}(t_k < y_i < t_{k+1} | t_k, t_{k+1})^{\mathrm{I}(w_i=k+1)} \tag{3.18}$$

for $0 < t_2 < \ldots t_{K-1} < t_{max}$ and 0 otherwise. As a function of t_k, this leads to the uniform process

$$p(t_k|\star) = \frac{1}{\min(y_i|w_i = k+1) - \max(y_i|w_i = k)}. \tag{3.19}$$

As depicted in Figure 3.1, the augmentation of the latent variable is an additional module in the hierarchical model framework, and hence the other parameters (except the residual variance that has been fixed to unity), and their fully conditional posterior densities, remain same as with the Gaussian response.

3.5.1 Binary response

Although the above threshold model is valid for a binary case-control response, the binary variables are often considered in a bit different manner. The binary response is usually coded as 0 and 1, instead of 1 and 2 as would be done in the above model when $K = 2$, and is thought to be linked to the Gaussian latent variable y_i such that

$$w_i = \begin{cases} 1, & \text{when} \quad y_i > 0 \\ 0, & \text{when} \quad y_i \leq 0. \end{cases}$$

Now the latent variable is given by the model equation (3.1) or (3.3) with a residual $\varepsilon_i \sim \mathrm{N}(0, 1)$, and hence the expected value of the binary variable becomes

$$\mathbb{E}(w_i) = \mathrm{P}(w_i = 1) = \mathrm{P}(y_i > 0) = \Phi(\mathbb{E}(y_i)).$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\mathbb{E}(y_i)$ being the linear predictor of the model (3.1) or (3.3). The probability of the binary variable, given the expected value of the latent variable, is Bernoulli with a success probability $\Phi(\mathbb{E}(y_i))$. This parametrization of the binary phenotype corresponds to a generalized linear model with the probit link function (Albert and Chib 1993).

3.5.2 Censored Gaussian response

A Gaussian phenotype with censored observations can be acquired *e.g.* as a logarithm of an event history or survival data, or as spiked phenotypes as *e.g.* in Broman (2003). For a censored Gaussian response we define an additional binary variable $\omega_i = 1$ if the ith observation is censored and $\omega_i = 0$ if not. As the threshold model assumes an unity variance for the latent Gaussian response, the observed phenotype must be standardized accordingly. This is done by regarding the available observations as a sample from a truncated normal density and utilizing the connection between the quantiles and the standard deviation of a Gaussian density. Now, if $\omega_i = 0$ the standardized Gaussian phenotype is used directly, and if $\omega_i = 1$ the underlying uncensored response is computed as previously. The latent variable parametrization of the censored phenotype corresponds to a generalized linear model with the tobit link function (see *e.g.* Tobin 1958; Sorensen *et al.* 1998; Iwata *et al.* 2009).

4 Parameter estimation

We perform the parameter estimation with a generalized expectation-maximization algorithm, that finds the maximum point of the joint posterior density by updating the parameters, one at the time, to the expected values of the above fully conditional posterior densities (3.4–3.14, 3.16 and 3.19). *Maximum a posteriori* or MAP estimate is the value that maximizes the posterior probability density function for the parameter vector Θ,

$$\widehat{\Theta}_{MAP} = \arg\max_{\Theta} \;\; p(data|\Theta)\, p(\Theta),$$

where $p(data|\Theta)$ denotes the likelihood of the data and $p(\Theta)$ the prior density. The MAP estimate differs from a maximum likelihood estimate

$$\widehat{\Theta}_{ML} = \arg\max_{\Theta} \;\; p(data|\Theta)$$

in that the MAP estimate incorporates the prior beliefs regarding the parameters values. Due to the conjugate or otherwise suitable prior densities chosen, the fully conditional posterior densities for the parameters and latent variables are known probability density functions. This guarantees an easy derivation of the estimation algorithm; as the expected values of the known densities are automatically available, we do not need to find the fully conditional posterior expectations by integration. Additionally, if preferred it would be trivial to implement a MCMC Gibbs sampler to sample from these densities.

4.1 Generalized expectation-maximization

The expectation-maximization algorithm is originally designed for imputation of missing data or estimation of latent variables, and it operates by iteratively updating the latent or missing variables to their expected values (E-step) and subsequently the parameter vector to its maximum likelihood value, given the values assigned to the latent variables (M-step) (Dempster *et al.* 1977). Later the algorithm has been used extensively in parameter estimation: in this case part of the parameter vector is updated to its expected value and the rest of the vector to its maximum likelihood value, or, in the Bayesian context, to the expected and maximum values of the fully conditional posterior densities, respectively. When an EM-algorithm is applied for parameter estimation, assigning the variables into the E- and M-steps is somewhat arbitrary. Often the parameters of most interest are maximized, while the variances and other nuisance parameters are integrated out from the posterior by updating them into their conditional expectations. As pointed out in I, under a Gaussian model the classification gets even more peculiar: due to the symmetric posterior density the expected and the maximum values of the location parameters are the same, and moreover the scale parameters with an inverse-χ^2 posterior became equivalent by a slight modification of the prior parameters. Thus, for it is not clear, or even interesting, which parameters are updated into their conditional maximums and which to conditional expectations, we base our method on an alternative description of the EM-algorithm (Neal and Hinton 1999) regarding both of the steps as maximization procedures of the same objective function. To enable handling of large marker sets the iterative updating is done one parameter at the time, conditionally on the other parameters remaining fixed. This practice is a form of a generalized expectation-maximization (GEM). Under the alternative description of the EM-algorithm, the GEM-algorithm corresponds to seeking to increase the objective function instead of attempting to maximize it. That is, we do not guarantee that the chosen arguments maximize the objective function, but know that the value of the function will increase with every update. The generalized algorithm has been proven to converge into same estimate than the standard EM-algorithm, though possibly slower (Neal and Hinton 1999).

4.2 Prior selection in MAP estimation

There are clearly two philosophies concerning the selection of the data specific (hyper)prior parameters, a theoretical one and a more pragmatic one. In theory, the parameter values could be elicited on the basis of the prior knowledge or beliefs, however, often it is more practical to choose values leading to the best result. The latter approach is referred as tuning of the estimation algorithm. Some authors disapprove this type of a prior selection, and suggest that it would be better to settle for a suboptimal model in order to be able to use philosophically more plausible priors (O'Hara and Sillanpää 2009). As in practice an end user scarcely will be content with an inferior model, we have in this case put the model performance before the philosophical affairs. Moreover, it is not clear what issues should be considered when trying to define the parameter values. The plausible values depend at least on the variance of the phenotypic response, on the heritability and genetic architecture of the trait, on the number of markers and on the LD-structure. Also the goal of the analysis must be taken into account when selecting the (hyper)prior parameters; within the QTL mapping context the most desirable result consists of clear and distinct QTL signals, with as little extra noise as possible, while in the prediction context this extra noise actually improves the model performance. There are attempts to determine the prior parameter values analytically (see *e.g.* Meuwissen *et al.* 2001; Shepherd *et al.* 2010; de los Campos *et al.* 2013), however, as these methods take into account only the phenotypic variance, heritability, and number of markers, they do not, in our experience, provide optimal results.

The information an EM-algorithm (or a GEM-algorithm) passes from one iteration to the next consists of one point of the posterior density, the maximum or the expectation. While in sampling based MCMC estimation the shape of the prior density is of crucial importance, in MAP estimation one needs to be more concerned about the behavior of the expectation or maximum points than the actual shape of the function. As our GEM-algorithm updates the parameters to their fully conditional expectations, we have mainly focused on the expected values in selecting the prior or hyperprior parameters. Since the tuning of the model will be the harder the more parameters there are to adjust, it is reasonable to try to manage with as few as possible. To this end, we have decided to set constant values to the (hyper)prior parameters having only a minor impact to the corresponding fully conditional expectations.

Under the Bayesian LASSO, or the Laplace model (I and III), the optional hyperprior layer is present in the model, and the the rate parameter λ^2 of the exponential density (a.k.a. the LASSO parameter) has a conjugate Gamma(κ, ξ) hyperprior. The conditional posterior expectation of the LASSO parameter is $\mathbb{E}(\lambda^2 \mid \star) = (\kappa + p)/(\xi + \sum \sigma_j^2/2)$; since p is very large, the impact of κ into the posterior expectation is negligible, and therefore the shape parameter κ of the gamma density is set to one. As we do not include the indicator variable into the Laplace model, the only parameter requiring tuning is the rate ξ of the gamma density. The shrinkage induced by the Laplace prior is the stronger the larger value the LASSO parameter gets. As the parameter ξ appears in the denominator of the posterior expectation, a low value for ξ is concordant with a high value for λ^2, and thus with more intense shrinkage. In II the prior parameters of the Extended Bayesian LASSO (Mutshinda and Sillanpää 2010) are handled correspondingly.

Under the Student's t model (I), the predetermined prior parameters for the inverse-χ^2 distribution are selected so that the fully conditional posterior expectation of the effect variance, $\mathbb{E}(\sigma_j^2 \mid \star) = (\beta_j^2 + \nu\tau^2)/(\nu - 1)$, is mainly determined by the the square of the effect size β_j. By setting the degrees of freedom $\nu = 2$, the posterior expectation will become $\beta_j^2 + 2\tau^2$, so if we choose a small value for the scale τ^2 the estimate of the variance stays always positive, but is shrunken towards zero strongly if β_j is small ($\beta_j << 1 \Longrightarrow \beta_j^2 << \beta_j$) while left intact when β_j is large ($\beta_j \approx 1 \Longrightarrow \beta_j^2 \approx \beta_j$). The scale parameter τ^2 is tuned into a data specific value. Contrary to the Bayesian LASSO, the Student's t model benefits from of the additional sparseness produced by the indicator variable. If the indicator variable is present in the model, the parameters requiring tuning are the scale τ^2 of the inverse-χ^2 density and the prior probability π of a marker to be linked to the trait. As these parameters are related to the shrinkage inducing mechanisms complementary to each other, their optimal values are interdependent. The value for π indicates the prior probability of a marker to be included into the model, so a low value naturally leads to stronger shrinkage than a high value. The scale parameter τ^2 behaves alike, low values producing more shrinkage.

The variance of the polygenic component (I and II), or the additive genetic variance under the Bayesian G-BLUP (I and III), has an inverse-χ^2 prior density. Similarly to the Student's t model, the degrees of freedom of the inverse-χ^2 density is set to $\nu_u = 2$ and the scale parameter τ_u^2 has been given a data specific value. The magnitude of the value given for τ_u^2 depends above all upon the role of the polygenic component in the model. When used as an additional polygenic component incorporated into a multilocus association model, the polygene is to explain only a fraction of the genetic variance, while in the context of Bayesian G-BLUP the polygene needs to cover the genetic variance in its entirety.

The (hyper)prior parameters are tuned into suitable data specific values by examining the model performance with different values, and selecting the ones leading to the most favorable outcome. What outcome is favorable depends naturally on the purpose of the analysis. In prediction the (hyper)prior parameters yielding the highest correlation between the predicted and observed phenotypes are selected, however *e.g.* in QTL mapping one may wish to use some other measure. In practice the optimization is done by simply selecting two arbitrary values for the parameter, observing the result under these values, and proceeding the search for an optimal value to the direction pointed by the better performing one. This step could be automatized, but so far we have performed it manually.

4.3 GEM-algorithm for a MAP estimate

The original papers I–III present slightly varying versions of the estimation algorithm, each corresponding to the model variate(s) considered in that particular paper. The algorithm presented here is a less-detailed summary of the previous versions.

1. Initial values

   Set initial values for the estimated (hyper)parameters. We use zeros for the location parameters β_0, $\boldsymbol{\beta}$ and $\mathbf{u}$, small positive values (0.1) for the dispersion parameters $\sigma_0^2, \boldsymbol{\sigma}^2, \sigma_u^2$ and λ^2 (δ and η_j under the EBL of II) and 0.5 for the indicators $\boldsymbol{\gamma}$. The possible unknown threshold values are initialized with $\mathbf{t}^\star = (\frac{1}{K-2}, \frac{2}{K-2}, \ldots, \frac{K-2}{K-2})$, where K is the number of classes. Some authors report sensitivity for starting values (*e.g.* Shepherd *et al.* 2010), but as we have not noticed such a behavior in our algorithm we are able to use always the same initial values.

2. Threshold model

   When the Gaussian phenotype is not (fully) observed, the threshold module is present the model, and the Gaussian response $\mathbf{y}$ is considered as a latent variable.

2.a. The values of the latent variable $\mathbf{y}$ are updated by replacing the current values of y_i with the expected values of the truncated normal distribution (3.16),

$$\mathbb{E}(y_i|\star) = \mathbb{E}(y_i) + \frac{\phi\big(t_{k-1} - \mathbb{E}(y_i)\big) - \phi\big(t_k - \mathbb{E}(y_i)\big)}{\Phi\big(t_k - \mathbb{E}(y_i)\big) - \Phi\big(t_{k-1} - \mathbb{E}(y_i)\big)},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions, respectively, and $\mathbb{E}(y_i)$ is the appropriate (given the sub-model) prior expectation for the latent response.

2.b. When applicable, the $K-2$ unknown thresholds are updated to their conditional expectations. From (3.19) we get

$$\mathbb{E}(t_k|\star) = \frac{1}{2}\Big(\max(y_i|w_i = k) + \min(y_i|w_i = k+1)\Big)$$

for all $k = 2, \ldots, K-1$.

3. Population intercept

The population intercept β_0 is updated into an appropriate version of the fully conditional expectation given by (3.4), the version referring to the different sub-models. The fully conditional posterior is a normal density, so $\mathbb{E}(\beta_0|\star)$ is given by its location parameter.

4. Residual variance

When the Gaussian response is fully observed, the residual variance σ_0^2 is updated into the expected value of an appropriate version of (3.5). The expected value of an Inv-χ^2(df, scale) density is (df × scale) / (df − 2), hence

$$\mathbb{E}(\sigma_0^2|\star) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \beta_0 - \ldots)^2.$$

5. Marker effect

5.a. The effect sizes β_j (for all j) are updated, one at the time, into an appropriate version of the fully conditional expectation given by the normal density in (3.6).

5.b. If the indicator variable is included into the model, the values of γ_j (for all j) are updated, one at the time, into the expected value of the Bernoulli distribution in (3.8)

$$\mathbb{E}(\gamma_j|\star) = p(\gamma_j = 1|\star) = \frac{\pi\mathcal{R}_j}{(1-\pi) + \pi\mathcal{R}_j},$$

where $\mathcal{R}_j$ is computed with an appropriate version of the formula given in (3.8).

Under the G-BLUP the marker effects are absent from the model, and this step becomes obsolete.

6. Polygenic component

   If the polygenic effect is included into the model, either as an additional component of the multilocus association model (3.1) or as the explanatory variable of the G-BLUP (3.3), the following updates will be carried out.

   6.a. The polygenic effect $\mathbf{u}$ is maximized by replacing the current value with the expected value of an appropriate version of the multivariate normal density (3.7).

   6.b. The additive variance of the polygenes σ_u^2 is replaced by its fully conditional expectation given by (3.11). As the fully conditional posterior is an Inv-χ^2 density, the expected value is

   $$\mathbb{E}(\sigma_u^2|\star) = \frac{1}{\nu_u + N - 2}\Big(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u\tau_u^2\Big) = \frac{1}{N}\Big(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + 2\tau^2\Big)$$

   for preset degrees of freedom $\nu_u = 2$. Under the G-BLUP the pedigree based relationship matrix $\mathbf{A}$ is replaced by its genomic counterpart $\mathbf{G}$.

7. Variance of the marker effects

   The effect variances σ_j^2 (for all j) are updated into their fully conditional expectations.

   7.a. Under the Student's t model the fully conditional posterior distribution of σ_j^2 is an inverse-χ^2, as expressed in (3.9), hence we get

   $$\mathbb{E}(\sigma_j^2|\star) = \frac{\beta_j^2 + \nu\tau^2}{\nu - 1} = \beta_j^2 + 2\tau^2, \text{ for preset degrees of freedom } \nu = 2.$$

   7.b. Under the Bayesian LASSO the precision, or inverse of the variance σ_j^2, has an inverse-Gaussian fully conditional posterior distribution (3.10) whose expected value equals

   $$\sigma_j^2 := \frac{|\beta_j|}{\lambda}.$$

7.c. Under the Indicator model in II the effects β_j have been given a Gaussian prior with a constant variance $\sigma_j^2 = \sigma^2$ for all j, and the effect variance is therefore not updated.

8. Prior parameters

   If the optional hyperprior layer is present in the model and update the values of λ and π into the fully conditional expectations.

   8.a. The expected value of a Gamma(shape, rate) density is shape/rate, hence, from (3.14) we get

$$\mathbb{E}(\lambda^2|\star) = \Big(1+p\Big)\Big(\xi + \sum_{j=1}^{p} \frac{\sigma_j^2}{2}\Big)^{-1} \text{ for preset shape } \kappa = 1.$$

   8.b. The expected value of a Beta(a, b) density is $a/(a+b)$, so, from (3.12):

$$\mathbb{E}(\pi|\star) = \frac{1}{2+p}\Big(1 + \sum_{j=1}^{p} \gamma_j\Big) \text{ for } a = b = 1,$$

   and

$$\mathbb{E}(\pi|\star) = \frac{1}{2p}\Big(a + \sum_{j=1}^{p} \gamma_j\Big) \text{ for } b = p - a,$$

   where p is the number of SNP markers, and a can be considered as the number of markers linked to the trait.

   In practice, while the LASSO parameter λ^2 is estimated in all of our algorithms in I–III, the estimation of the probability π is seldom carried out. The indicator variable is either absent from the model (as the hierarchical Laplace model does not need it), or the prior value for π is fixed (Student's t model in I and the Indicator model in II).

The steps are repeated until convergence.

5 Example analyses

In the original works I–III we have examined the behavior and performance of the different sub-models, and tested and demonstrated our method in both prediction and association mapping context. During the analyses we have observed especially the prediction (I and III) and association mapping (II) accuracy, and the ease of finding the suitable (hyper)prior parameters.

We have used three different data sets, consisting both simulated and real data. The data sets represent various population structures, genetic architectures and linkage disequilibrium patterns.

5.1 Data sets

5.1.1 XIII QTL-MAS Workshop data

First of the data sets, used in all of the original works I–III, consists of a simulated data introduced in the XII QTL-MAS Workshop 2008 (Lund *et al.* 2009). The data set can be downloaded from the workshop homepage, `http://www.computationalgenetics.se/ QTLMAS08/QTLMAS/DATA.html`. This is a extensively-used data (*e.g* Usai *et al.* 2009; Hallander *et al.* 2010; Shepherd *et al.* 2010), and therefore enables an easy way to get some idea of the performance of our method in comparison to other methods proposed. The data set consists the genotypes of 6,000 biallelic SNP loci of 5,865 individuals from seven generations of half sib families, simulating a typical livestock breeding population (see Lund *et al.* 2009 for details). The first four generations of the data, 4,665 individuals, function as a learning set, while the generations five to seven, 1,200 individuals, are treated as a prediction set. The advantage of using a simulated data set in the example analyses is the availability of the true genetic values of the individuals, and the true effects and locations of the simulated causal loci. The individuals' genetic value equals a cumulative effect of 48 simulated QTL, and the phenotypic values have been obtained as a sum of the genetic value and a random residual drawn from a normal density with null mean and a variance set to produce heritability value 0.3 (Lund *et al.* 2009). To examine the model performance in a less data specific situation, with the influence of sampling variation diminished, we have generated 100 replicates of the data set by resampling the residuals from a normal density $\mathrm{N}(0, var(TBV)(1/h^2 - 1))$, where $var(TBV)$ denotes the observed variance of the genetic values and the heritability h^2 equals 0.3.

As in II our main interest was the behavior of the model with and without an additional polygenic component, we modified the data by adding a simulated polygenic effect into the phenotype. We simulated the polygenic effect by either sampling the polygene from a pedigree based multivariate normal density, or by selecting 1000 random SNP to serve as codominant causal loci with equal allele substitution effects. The former method corresponds to the Fisher's polygenic model, while the latter can be seen a finite locus approximation of the polygenic model. When generating the 100 replicated phenotypes, aside of the residuals, the infinite polygenic terms and the random SNP acting as the polygene were resampled.

5.1.2 Real pig (*Sus scrofa*) data

The second data set is a real pig (*Sus scrofa*) data that we have used in I and III. The data is provided by the Genetics Society of America to be used for benchmarking of genomic selection methods, and it is described in detail by Cleveland *et al.* (2012). The pig data set consists of phenotypic records of 3,184 individuals for trait with predetermined heritability 0.62, and genotypic records for 60k biallelic SNP markers. Since the data does not consist a separate validation population, we compute the result statistics using cross-validation, where the 3,184 individuals are randomly partitioned into 10 subsets (10-fold cross-validation) of 318 or 319 individuals. At each round 9 of the sets are treated as a learning set and the remaining one as the prediction set.

5.1.3 Human HapMap data

In the third data set, used in the original work II, the SNP genotypes come from the International HapMap Project (The International HapMap Consortium 2003) phase 3 data, available at `http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/`. The data consists of 1184 individuals from 11 populations around the world. For our data set we selected the SNP loci from chromosome 1, that have no missing genotypes and minor allele frequency (MAF) more than or equal to 0.05, leading to a set of 31,916 markers. The phenotype data was created by selecting 10 random markers with MAF > 0.4 as QTL and drawing the allele substitution effects of the QTL from a Gamma(4,0.5) density. The MAF limit was assigned to produce QTL detectable with the limited number of individuals. As there is no pedigree available in this data set, the finite locus approximation approach was the only possibility for the polygene simulation, and so the polygenic effect was added by randomly selecting 500 markers as a polygenic loci with equal allele substitution effects. The heritability was set to 0.5, and 80% of the additive genetic variation was set to be due to the QTL and 20% due to the polygene. To create a realistic situation where the SNP are not causal mutations all of the simulated causal loci (total of 510 markers) were removed from the marker data. A set of 100 phenotype data replicates was created by resampling the residuals.

5.1.4 Discrete and censored data

The binary, ordinal and censored data sets used in I (only binary) and III are based on the QTL-MAS (in I and III) and the real pig (III only) data sets. The binary and ordinal data were generated by discretizing the original phenotypes by introducing one or several thresholds at selected positions. The binary response in I is set to have 80% success probability, while in III there are two binary phenotypes with success probabilities 50% and 80%. The ordinal phenotypes consists of four classes with either even 20:30:30:20% of observations in each class, or highly unbalanced with 70% belonging to the first class and 10% in the subsequent three classes. The censored data sets consist of a continuous Gaussian phenotype with 20-, 50- or 80% right censored observations. The value of the censored observations is set to equal the largest of the non-censored values, leading to a spiked Gaussian phenotype (see Broman 2003). The binary and the evenly distributed ordinal data sets are generated in preparation for an easy ascertainment of the the extra power acquired by utilizing the category information compared to the dichotomized phenotype. The binary phenotype with 80% success probability simply sets the first category of the ordinal phenotype as a failure and the subsequent three classes as a success while the binary response with 50% success probability sets the first and second category as a failure and the third and fourth as a success. The same holds true for the censored data, as the threshold values are set to correspond the thresholds of the binary phenotype. All threshold values were determined as standard normal distribution function parameters leading to the desired threshold value, *e.g.* a threshold at 0.84 leading to 20% success probability, since $\Phi(0.84) = 0.8$.

5.2 Pre-selection of the markers

The multilocus models are not able to handle an unlimited number of loci with respect to the sample size. In the QTL-MAS data set (I–III) the proportion of markers to individuals is almost a one-to-one, and no extra measures are needed, but with the pig (I and III) and the HapMap (II) data sets the multilocus association model becomes too oversaturated to function properly. Therefore, prior to the association analysis, we have reduced the number of markers by applying the sure independence screening method of Fan and Lv (2008) for ultrahigh dimensional feature space. The sure independence screening is based on ranking the predictors with respect to their marginal correlation with the response variable, and selecting either a predetermined proportion of the predictors or the predictors exceeding a predetermined importance measure.

Hoti and Sillanpää (2006) have proposed an upper limit of 10 times more loci than individuals, but it seems that in practice a smaller number of loci might be optimal. In I and III we have reduced the number of markers in the pig data from 45,317 to 10,000 and in II in the HapMap data from 31,916 to 5,000 markers, leading to roughly three and four times more markers than individuals, respectively.

5.3 Genomic prediction

The performance of the model from the prediction aspect has been considered in original works I and III. In I we compared the predictive performance of the model with the two alternative shrinkage priors, Student's *t* and Laplace, and studied the impact of the model components into the accuracy of the estimates. In III we tested and demonstrated the behavior of the threshold model for binary, ordinal and censored Gaussian traits.

To compare the accuracy of the estimates we computed the genomic breeding values for the prediction set individuals and examined the Pearson's product-moment correlation coefficient between the true and the estimated breeding values under the model variants. In the simulated QTL-MAS data set the genetic values of the individuals are known, enabling us to determine the accuracy by a direct comparison of the simulated and estimated genetic values. With a real data, on the other hand, the true breeding values of the individuals are not available, and therefore the correlation for the pig data is computed between the estimated breeding values and the phenotype, and divided by the square root of the heritability to compensate for the additional noise. For this we have used the heritability value given in Cleveland *et al.* (2012). In both cases the estimated breeding values are simply computed as the linear predictor of the current model.

5.4 Association mapping

The association mapping perspective of the method is considered in the original work II. Association mapping aims to locate genes affecting the phenotype by identifying the SNP markers in linkage disequilibrium with the causal loci.

5.4.1 Decision making

An irremovable part of association mapping consists of deciding which observed SNP effects should be considered as a signal for a causal locus and which should be ignored as a random noise. The decision making poses a two fold problem: one hopes to find as many of the real associations as possible, while avoiding the false signals. While the amount of false signals, or false positive rate, one is willing to accept is decided beforehand by setting a suitable confidence level for the experiment, the signal size corresponding to that level is case-specific. Phenotype permutation (Churchill and Doerge 1994; Xu 2003) is an universal, although computer intensive, method for assessing empirical confidence limits. The significant SNP effects are identified by randomly shuffling the phenotypes T times, recording the highest SNP effect of each permutation round, and considering the tth highest effect as the $(T - t/T)\%$ confidence limit. The SNP effects higher than the confidence limit are then judged as genuine signals.

Under the Indicator model in II also the Bayes factor can be used in validating the signals. As the normal prior for the effect sizes β_j does not introduce shrinkage to the estimates, the posterior expectation of the indicator variable truly represents the probability of the marker to be linked to the trait. Hence the Bayes factor can be defined as the ratio of the marginal likelihoods of the two models, the first model corresponding to indicator $\gamma_j = 1$ (SNP is linked to the trait) and the second model to indicator $\gamma_j = 0$ (SNP is not linked to the trait). Since in an Indicator model like ours there exists only the two competing models, the Bayes factor is often computed as the posterior odds of the models divided by the prior odds of the models, that is, the Bayes factor related to marker locus j is simply given by

$$\mathrm{BF}_j = \frac{\hat{\gamma}_j}{1 - \hat{\gamma}_j} \Big/ \frac{\pi}{1 - \pi},$$

where $\hat{\gamma}_j$ is the posterior estimate of the indicator (posterior probability that marker j is linked to a QTL) and π the prior probability that marker is linked to a QTL.

5.4.2 Diagnostics

The model performance was assessed by examining the numbers of true (n_{tp}) and false (n_{fp}) positive, as well as true (n_{tn}) and false (n_{fn}) negative signals, and computing the false positive (FPR), false negative (FNR) and false discovery (FDR) rates as

$$\text{FPR} = \frac{n_{fp}}{n_{fp} + n_{tn}}, \quad \text{FNR} = \frac{n_{fn}}{n_{tp} + n_{fn}} \quad \text{and} \quad \text{FDR} = \frac{n_{fp}}{n_{fp} + n_{tp}}.$$

The FPR and FNR correspond to the probabilities of a type I error (1− specificity) and a type II error (or 1− power), respectively, and FDR represents the proportion of the reported QTL that are actually false positives. A method was considered to have correctly identified a QTL if it reported one or more QTL signals within a window of predetermined width around a known (simulated) QTL. The number of true positives (n_{tp}) was the number of windows consisting one or more signals. Respectively, the number of false negatives (n_{fn}) was the number of windows without a QTL signal. Reported QTL outside the windows were treated as false positives (the number of false positives being denoted by n_{fp}). The number of true negatives (n_{tn}) was calculated as the number of the SNP outside the windows around the simulated QTL minus the number of false positives.

To compare the performance of the different models, we examined the average true positive rates (TPR = 1−FNR, or sensitivity of the method) against the false detection rates within analyses of the 100 replicated data sets, under a series of limit values for the SNP effects considered as a positive signal. The most common diagnostic graph, the ROC-curve, (Receiver Operating Characteristic curve), plots the true positive rate (TPR) against the false positive rate (FPR), however we feel that replacing the latter with the false discovery rate (FDR) leads to a more intuitive presentation when the number of markers is high (Figure 1 in II). As the number of segregating QTL is usually negligible compared to the number of SNP markers in a genome-wide data, and hence the number of false positives and the number of true negatives approximately add up to the number of SNP, FPR measures approximately the proportion of the markers giving a false signal. Given that there are tens or hundreds of thousands of markers, a substantially low percentage of false positives will vastly exceed the number of true positives, and lead to a situation where most of the validated signals are in fact false. The false discovery rate, on the other hand, represents the proportion of the reported QTL that are actually false positives, and therefore tells exactly what one is going to get.

5.5 Of speed and convergence

The steps of the GEM-algorithm are repeated until convergence. The algorithm is considered to be converged when the correlation between the estimated breeding values of two consecutive iterations is higher than $1-10^{-6}$. The convergence is confirmed visually, by examining the behavior of parameter values during the iterations, and verifying that all of the parameters have reached a constant level. This is also how the suitable value for the convergence rule has been originally ascertained. The required number of iterations is usually between 40 and 80 under the multilocus association model, and around 10 under the G-BLUP. So far we have not encountered problems in the convergence, given that appropriate hyper(prior) parameter values have been selected, and that the number of markers with respect to the sample size has not been too large.

Depending on the data and the model variate, the computation time is around 10–40 seconds. As a MCMC algorithm usually takes hours at minimum to converge, the speed difference between the two types of algorithms is far from trivial. The extremely short time requirement in fact enables the usage of computer intensive techniques such as phenotype permutation.

6 Conclusions

6.1 Current status

Genomic selection has proven to accelerate the genetic gain of a breeding program compared to phenotypic selection (Schaeffer 2006). While the power of the traditional marker assisted selection is negligible when the trait is controlled by a large number of small QTL whose effects can not be reliably identified, in genomic selection this problem is largely avoided by passing the decision making between QTL and non-QTL and including all of the effects into the genomic breeding value estimate. In the animal, especially cattle, breeding field genomic selection is widely accepted as the new paradigm, and genomic breeding values are used in national and international cattle breeding programs in several countries (Zhang *et al.* 2011; Eggen 2012). The plant breeding community, on the other hand, is still investigating the practical value of genomic selection (see *e.g.* Jannink *et al.* 2010; Nakaya and Isobe 2012).

In human genetics the genome-wide association (GWA) studies have revealed hundreds of validated associations between SNP markers and complex traits (see *e.g.* Donnelly 2008). However, for any one trait the validated associations typically explain only a fraction of the observed genetic variation, causing the so called missing-heritability-problem (Maher 2008).

Many of the suggested explanations for the missing heritability (epistasis, gene-environment interaction, epigenetic factors), are based on non-additive genetic effects, and hence are discarded as possible explanations for the missing additive genetic variance or narrow-sense heritability (de los Campos *et al.* 2010; Yang *et al.* 2010). By using a frequentist G-BLUP, Yang *et al.* (2010) has found a remarkable part of the missing heritability of human height: while validated SNPs explain 5% of the phenotypic variance, the G-BLUP explains 45%, indicating that the GWA analyses have detected only a proportion of the QTL. This result suggests two things. First, the GWA study has not been efficient enough, as it has failed to detect many of the small causal loci, probably because the small effects do not reach the significance thresholds and are discarded as random noise. This problem can be overcome by increasing the power of the GWA by increasing sample size, marker density and statistical methodology. However, it is probable that many of the causal loci explain such a small amount of the phenotypic variation that they will never be detected, and on the other hand, expending considerable resources in detecting practically insignificant loci may not be the best policy. This brings us to the second implication. As the breeding field has witnessed with the genomic selection, in prediction it is not necessary to know the exact location and effect of the specific causal variants, but only the total effect (Meuwissen *et al.* 2001). Therefore, it might be well advised to separate the gene-detecting GWA studies from the phenotype and disease risk prediction (see *e.g.* de los Campos *et al.* 2010; Yang *et al.* 2010).

The accuracy of genomic breeding values estimated for a given species for a given trait depends at least on the effective population size and the genome length of the species, the heritability and the genetic architecture of the trait, the size and structure of the training set, the density of the marker map and the statistical approach used for estimating the genomic breeding values (Zhang *et al.* 2011). As the properties of the species and the trait can not be altered, the improvement must be acquired via the training set selection and the density of the markers, and by developing better statistical methodology. Muir (2007) has shown that the size and structure (number of generations) of the training set affects the accuracy more than the number and density of the markers. As noted earlier, the optimal number of markers in the multilocus association model is a function of the number of individuals in the model, so increasing the size of the training set may improve the outcome also through this channel. Different statistical methods of predicting the genomic estimated breeding values have been compared in several studies (see *e.g.* Calus and Veerkamp 2007; Calus 2010; Moser *et al.* 2009; Daetwyler *et al.* 2010), and it seems that there is no single method working best at all situations.

6.2 What have we learned?

In the original works I and III we have further confirmed the idea of the mutually complementary nature of the Bayesian multilocus association model and the G-BLUP in the genomic selection. When the trait is influenced by a moderate number of causal genes (tens, but not hundreds), the multilocus association model seems to give substantially higher accuracy for the predictions than the Bayesian G-BLUP. With a highly polygenic trait the Bayesian G-BLUP may, however, have the advantage. As in all of the analyses our Bayesian version of the G-BLUP has been able to predict the breeding values without any prior knowledge about the variance components, it seems that the Bayesian G-BLUP may be a serious rival for ASREML and other frequentist BLUP-type methods.

On the basis of the original work I it seems that, at least in MAP-estimation context, the hierarchically formulated Laplace prior density is superior to the non-hierarchical Laplace density and to the hierarchically formulated Student's t prior density. Especially with a polygenic trait (the pig data) the hierarchical Laplace prior provides clearly more accurate genomic breeding value estimates. This is an important discovery, since the critique towards marker association models and their bad behavior in a polygenic situation (*e.g.* Daetwyler *et al.* 2010; Clark *et al.* 2011) is mainly based on the observations on BayesB, which is a Student's t model.

The original work II reinforced our prior conception of the robustness of the multilocus association model to residual dependencies between the individuals. Based on this work, we are confident to say that multilocus association methods improve the QTL mapping performance compared to single SNP methods. Further, to our experience, it seems that neither an additional correction for population or sample structure, nor an additional polygenic component, is required under a multilocus association method. In II we hypothesized that the shrinkage based multilocus association model *per se* might provide a possible explanation for the apparent redundancy of the polygenic component. As the shrinkage based multilocus association models incorporate all of the marker effects, not only the substantial ones, the negligible-sized SNP effects may contribute to the model as a finite locus approximation of the polygenic component. In a high density genetic map the approximation may be good enough to leave a separate polygenic component obsolete. Similarly, if some loci within the marker set were correlated with the population or ethnic memberships, in a high density map these loci could serve as an approximation of an explicit population or ethnic term (Wang *et al.* 2005).

In addition, our example analyses in II demonstrate that Bayesian multilocus association approaches can improve QTL mapping accuracy and avoid occurrences of biased association signals due to model misspecification. In the excample analyses the uncorrected single SNP model PLINK (Purcell *et al.* 2007) found basically nothing, while the multilocus association model was in most cases able to identify the QTL explaining more than 1% of the phenotypic variance.

In the original work III we proposed the threshold model part of the model framework. On the basis of our findings it is unclear whether the additional latent parameter module actually improves the prediction accuracy compared to using the linear Gaussian model directly for the binary, ordinal or censored response. However, even though our results do not confirm the practical superiority of the correct threshold model over the linear Gaussian model, we urge caution when applying a Gaussian model directly for an ordinal data. Some data sets may be less well-behaving than the ones we have studied and, as proven by Wang *et al.* (2013), different linear models may be less robust to the incompatible data.

In all of the original works we have reduced the number of markers in the multilocus association model by preselecting the markers with the sure independence screening (Fan and Lv 2008). Even though sure independence screening is a strikingly simple method it works very well, probably because all it needs to do is to let all of the important markers pass to the next step while, since the final variable regularization is performed by the multilocus association model, it does not matter if unimportant ones are also selected.

6.3 What's next?

So far we have built the genomic relationship matrix used in the Bayesian G-BLUP with the simple identity-by-state method proposed by VanRaden (2008). As it seems to be clear that G-BLUP is a worthy method with a highly polygenic trait, it would probably be advised to try and develop a more realistic method for assessing the relationships, possibly by trying to take the dependencies between the markers into account.

The marker preselection procedure needs more attention. While in all of the original works we have used the sure independence screening, on the side we have also performed some preliminary testing with other methods. Even though the sure independence screening works actually surprisingly well compared to the more sophisticated methods we have tested, there still may be room for improvement.

The human genetic field intrigues us for a couple of reasons. As the human genetics field is in possession of the most extensive data sets, it would be interesting to test the performance of Bayesian multilocus association model framework in that context. Also, it seems that the statistical genome-wide association analysis methods typically used in the human genetics field may not be as efficient as they could be. It would therefore be especially fascinating to test our hypothesis of the superiority of the Bayesian multilocus association models or, when applicable, the Bayesian G-BLUP, compared to the predictive and gene-detection abilities of the traditional methods.

Acknowledgments

This work was financially supported by research grants from the Academy of Finland, the University of Helsinki's Research Funds and Tekes funding for the public research "Individualized Connected Health".

References

Albert, J. H. and S. Chib, 1993 Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88:** 669–679.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. Van Duijn, 2007 GenABEL: an R library for genome-wide association analysis. Bioinformatics **23:** 1294–1296.

Ayers, K. L. and H. J. Cordell, 2010 SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet. Epidemiol. **34:** 879–891.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevems, Y. Ramdoss, and E. S. Buckler, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633–2635.

Broman, K. W., 2003 Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics **163:** 1169–1175.

Calus, M. P. L., 2010 Genomic breeding value prediction: methods and procedures. Animal **4:** 157–164.

Calus, M. P. L. and R. F. Veerkamp, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J. Anim. Breed. Genet. **124:** 362–368.

Chhikara, R. and L. Folks, 1989 *The inverse Gaussian distribution: theory, methodology, and applications*. New York, NY, USA: Marcel Dekker, Inc.

Cho, S., K. Kim, Y. J. Kim, J. K. Lee, Y. S. Cho, J. Y. Lee, B. G. Han, H. K. ans J. Ott, and T. Park, 2010 Joint identification of multiple genetic variants via elastic-net variable selection in a genome wide association analysis. Ann. Hum. Genet. **74:** 416–428.

Churchill, G. A. and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Clark, S. A., J. M. Hickey, and J. H. J. van der Werf, 2011 Different models of genetic variation and their effect on genomic evaluation. Genet. Sel. Evol. **48:** 18.

Cleveland, M. A., J. M. Hickey, and S. Forni, 2012 A common dataset for genomic analysis of livestock populations. G3 **2:** 429–435.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics **185:** 1021–1031.

de los Campos, G., D. Gianola, and D. B. Allison, 2010 Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. **11:** 880–886.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193:** 327–345.

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

Dellaportas, P., J. J. Forster, and I. Ntzoufras, 2002 On Bayesian model and variable selection using MCMC. Stat. Comput. **12:** 27–36.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39:** 1–38.

Donnelly, P., 2008 Progress and challenges in genome-wide association studies in humans. Nature **456:** 728–731.

Eggen, A., 2012 The development and application of genomic selection as a new breeding paradigm. Anim. Front. **2:** 10–15.

Ewans, W. J. and R. S. Spielman, 1995 The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. **57:** 455–464.

Fan, J. and J. Lv, 2008 Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. B **70:** 849–911.

Fang, M., D. Jiang, D. Li, R. Yang, W. Fu, L. Pu, H. Gao, G. Wang, and L. Yu, 2012 Improved LASSO priors for shrinkage quantitative trait loci mapping. Theor. Appl. Genet. **124:** 1315–1324.

Figueiredo, M. A. T., 2003 Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. **25:** 1150–1159.

Gelman, A., 2004 Parameterization and Bayesian modeling. J. Am. Stat. Assoc. **99:** 537–545.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004 *Bayesian Data Analysis* (2nd ed.). USA: Chapman & Hall / CRC.

George, E. I. and R. E. McCulloch, 1993 Variable selection via gibbs sampling. J. Am. Stat. Assoc. **88:** 881–889.

Gilks, W., S. Richardson, and D. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2009 *AS-Reml User Guide Release 3.0.* Hemel Hempstead, UK: VSN International Ltd,.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics **12:** 186.

Hallander, J., P. Waldmann, C. Wang, and M. J. Sillanpää, 2010 Bayesian inference of genetic parameters based on conditional decompositions of multivariate normal distributions. Genetics **185:** 645–654.

Hayashi, T. and H. Iwata, 2010 EM algorithm for Bayesian estimation of genomic breeding values. BMC Genetics **11:** 3.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics **31:** 423–447.

Hoerl, A. E., 1962 Application of ridge analysis to regression problems. Chem. Eng. Prog. **58:** 54–59.

Hoggart, C. J., J. C. Whittaker, M. De Iorio, and D. J. Balding, 2008 Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet. **4:** e1000130.

Hoti, F. and M. J. Sillanpää, 2006 Bayesian mapping of genotype × expression interactions on quantitative and qualitative traits. Heredity **97:** 4–18.

Iwata, H., K. Ebana, S. Fukuoka, J.-L. Jannink, and T. Hayashi, 2009 Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa L.* germplasms. Theor. Appl. Genet. **118:** 865–880.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics **9:** 166–177.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. E. Kong, and N. B. Freimer, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42:** 348–354.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, 2008 Efficient control of population structure in model organism association mapping. Genetics **178:** 1709–1723.

Kuo, L. and B. Mallick, 1998 Variable selection for regression models. Sankhya Ser. B **60:** 65–81.

Lange, K., 1997 *Mathematical and Statistical Methods for Genetic Analysis* (1st ed.). Statistics for Biology and Health. New York: Springer.

Lee, S. H., M. E. Goddard, P. M. Visscher, and J. H. van der Werf, 2010 Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genet. Sel. Evol. **42:** 22.

Lee, S. H., J. H. J. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, 2008 Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. **4:** e1000231.

Li, J., K. Das, G. Fu, R. Li, , and R. Wu, 2011 The Bayesian LASSO for genome-wide association studies. Bioinformatics **27:** 516–523.

Lund, M. S., G. Sahana, D.-J. de Koning, G. Su, and O. Carlborg, 2009 Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proc. **3:** S1.

Maher, B., 2008 Personal genomes: The case of the missing heritability. Nature **456:** 18–21.

McLachlan, G. J. and T. Krishnan, 1997 *The EM Algorithm and Extensions*. Hoboken, New Jersey: John Wiley & Sons, INC.

Meuwissen, T. H. E., B. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819–1829.

Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet. Sel. Evol. **41:** 2.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma, 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. **41:** 56.

Muir, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. **124:** 342–355.

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.

Nakaya, A. and S. N. Isobe, 2012 Will genomic selection be a practical method for plant breeding? Ann. Bot. **110:** 1303–1316.

Neal, R. M. and G. E. Hinton, 1999 A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*, pp. 355–368. Cambridge, MA, USA: MIT Press.

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

Park, T. and G. Casella, 2008 The Bayesian Lasso. J. Am. Stat. Assoc. **103:** 681–686.

Pikkuhookana, P. and M. J. Sillanpää, 2009 Correcting for relatedness in Bayesian models for genomic data association analysis. Heredity **103:** 223–237.

Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. **11:** 800–805.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. **81:** 559–575.

Schaeffer, L., 2006 Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet **123:** 218–223.

Shepherd, R. K., T. H. Meuwissen, and J. A. Woolliams, 2010 Genomic selection and complex trait prediction using a fast EM algorithm applied to genomewide markers. BMC Bioinformatics **11:** 529.

Sillanpää, M. J. and M. Bhattacharjee, 2005 Bayesian association-based fine mapping in small chromosomal segments. Genetics **169:** 427–439.

Sorensen, D. A., S. Andersen, D. Gianola, and I. Korsgaard, 1995 Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. **27:** 229–249.

Sorensen, D. A., D. Gianola, and I. R. Korsgaard, 1998 Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. Acta Agric. Scand. **48:** 222–229.

Sun, W., J. G. Ibrahim, and F. Zou, 2010 Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics **185:** 349–359.

The International HapMap Consortium, 2003 The International HapMap Project. Nature **426:** 789–796.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58:** 267–288.

Tobin, J., 1958 Estimation of relationships for limited dependent variables. Econometrica **26:** 24–36.

Usai, M. G., M. E. Goddard, and B. J. Hayes, 2009 LASSO with cross-validation for genomic selection. Genet. Res. **91:** 427–436.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. **91:** 4414–4423.

Verbyla, K. L., B. J. Hayes, P. J. Bowman, and M. E. Goddard, 2009 Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet. Res. **91:** 307–311.

Wang, C.-L., X.-D. Ding, J.-Y. Wang, J.-F. Liu, W.-X. Fu, Z. Zhang, Z.-J. Yin, and Q. Zhang, 2013 Bayesian methods for estimating GEBVs of threshold traits. Heredity **110:** 213–219.

Wang, Y., R. Localio, and T. R. Rebbeck, 2005 Bias correction with a single null marker for population stratification in candidate gene association studies. Hum. Hered. **59:** 165–175.

Weeks, D. and G. Lathrop, 1995 Polygenic disease: methods for mapping complex disease traits. Trends Genet. **11:** 513–519.

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

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

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

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. **42:** 565–571.

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

Yi, N., V. George, and D. B. Allison, 2003 Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics **164:** 1129–1138.

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

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38:** 202–208.

Zhang, Z., Q. Zhang, and X. Ding, 2011 Advances in genomic selection in domestic animals. Chi. Sci. Bull. **56:** 2655–2663.