

# Bayesian Linear Regression Models

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-03-10

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Bayesian Linear Regression Model for Marker Effects</b>	<b>1</b>
2.1	Statistical model and model parameters . . . . .	1
2.2	Prior distributions of model parameters . . . . .	2
2.3	Estimation of model parameters . . . . .	6
2.4	Mapping of variants . . . . .	7
2.5	Extensions to Multiple Traits and Multiple Components . . . . .	8
2.6	Extensions to Summary Statistics . . . . .	10
<b>3</b>	<b>Bayesian Linear Regression Model for Individual Effects</b>	<b>13</b>
3.1	Statistical model and model parameters . . . . .	13
3.2	Extensions to multiple components . . . . .	15
3.3	Estimation of model parameters . . . . .	16

## 1 Introduction

Bayesian linear regression models have been proposed as a unified framework for mapping, prediction, estimation of heritability and effect distribution (Moser et al. 2015).

Bayesian linear regression models attempts to account for the underlying genetic architecture of the trait. This is achieved by using many linked markers covering the entire genome to jointly estimate marker effects, and by allowing for the genetic signal to be heterogeneous distributed over the genome (i.e. some regions have stronger genetic signal than others). This may in some situations allow a more accurate estimate of the true underlying genetic signal leading to more accurate predictions.

Bayesian linear regression models can also be used to map variants associated with phenotypes and to estimate the total variance explained by the SNPs. Because they fit all SNPs simultaneously and account for LD between SNPs, they should have greater power to detect associations, find less false negatives and give unbiased estimates of the larger SNP effects. They can also provide information about the genetic architecture of the trait from the hyper-parameters of the distribution of SNP effects.

Bayesian linear regression models fit all markers simultaneously and their effects as drawn from a prior distribution that attempts to match the true distribution of SNP effects as closely as possible. However

the true distribution of effect sizes is unknown but a mixture of normal distributions can approximate a wide variety of distributions by varying the mixing proportions. Erbe et al. used this prior and included one component of the mixture with zero variance. A similar model was proposed by Zhou et al. but with a mixture of two normal distributions, one with a small variance and one with a larger variance.

In the following the statistical model, prior distributions, algorithms for estimation of model parameters, extensions for handling multiple marker sets and multiple traits for the Bayesian linear regression models is presented. We will present to alternative BLR models, BLR model for marker effects and BLR model for individual effects.

## 2 Bayesian Linear Regression Model for Marker Effects

### 2.1 Statistical model and model parameters

The phenotype is related to the set of genetic markers under the multiple linear regression model:

$$\mathbf{y} = \mathbf{W}\mathbf{b} + \mathbf{e} \quad (1)$$

where  $\mathbf{y}$  is the phenotype,  $\mathbf{W}$  a matrix of SNP genotypes, where values are standardised to give the  $ij$ th element as:  $w_{ij} = (x_{ij} - 2p_j) / \sqrt{2p_j(1 - p_j)}$ , with  $x_{ij}$  the number of minor (or alternative) alleles (0, 1 or 2) for the  $i$ th individual at the  $j$ th SNP and  $p_j$  the minor (or alternative) allele frequency.  $\mathbf{b}$  are the genetic effects for each SNP, and  $\mathbf{e}$  the residual error. The dimensions of  $\mathbf{y}$ ,  $\mathbf{W}$ ,  $\mathbf{b}$  and  $\mathbf{e}$  are dependent upon the number of traits,  $k$ , the number of SNP markers,  $m$ , and the number of individuals,  $n$ , and are described in the sections below.

The residuals,  $\mathbf{e}$ , are a priori assumed to be independently and identically distributed multivariate normal with null mean and covariance matrix  $\mathbf{I}\sigma_e^2$ .

### 2.2 Prior distributions of model parameters

Most complex traits and diseases are likely highly polygenic, with hundreds to thousands of causal variants, most frequently of small effect. So, the prior distribution must include many small and few large effects. Furthermore marker effects are a priori assumed to be uncorrelated (but markers can be in strong linkage disequilibrium and therefore a high posterior correlation). Many priors for marker effects have been proposed. These priors come more from practical (ease of computation) than from biological reasons. Each prior originates a method or family of methods, and we will describe them next, as well as their implications.

1. Normal distribution: Random regression BLUP (RR-BLUP), SNP-BLUP, G-BLUP
2. Normal distribution with unknown variances: BayesCPi0, GREML
3. Student (t) distribution : BayesA
4. Mixture of Student (t) distribution and spike at 0: BayesB
5. Mixture of Normal distributions and spike at 0: BayesCPi and BayesR
6. Double exponential: Bayesian Lasso

#### Prior marker variance SNP-BLUP (RR-BLUP and G-BLUP)

In SNP-BLUP the prior for each marker effect follows a priori a normal distribution with a variance  $\sigma_b^2$  which is constant across markers:

$$p(b) = \prod_i p(b_i) \quad (2)$$

where

$$p(b_i) = N(0, \sigma_b^2) \quad (3)$$

In a normal distribution most effects are concentrated around 0, whereas few effects will be large. Therefore, the prior assumption of normality precludes few markers of having very large effects – unless there is a lot of information to compensate for this prior information.

The allele frequencies should be the ones in the population where the genetic variance was estimated (e.g., the base population of the pedigree) and not the current, observed populations.

We assume that markers are independent one from each other. This can be equivalently written as:

$$p(b) = MVN(0, D) \quad (4)$$

$$Var(b) = D = I\sigma_b^2 \quad (5)$$

where MVN stands for multivariate normal. This formulation including D will be used again throughout these notes.

### 2.2.1 Prior marker variance

Marker “variances”, can, however, be included within a Bayesian framework. The Bayesian framework will postulate a non-normal distribution for marker effects, and this non-normal distribution can be explained as a two-stages (or hierarchical) distribution. In the first stage, we postulate that each marker has a priori a different variance from each other:

$$p(b_i | \sigma_{b_i}^2) = N(0, \sigma_{b_i}^2) \quad (6)$$

In the second stage, we postulate a prior distribution for the variance themselves:

$$p(\sigma_{b_i}^2 | something) = p(\dots) \quad (7)$$

This prior distribution helps (the estimate of  $\sigma_{b_i}^2$  is more accurate, in the sense of lower mean square error) although it will still be far from reality (e.g. (Gianola et al. 2009)). This is very convenient because the solving algorithm simplifies greatly and most Bayesian regression models are based in this idea.

### Prior marker variance Bayes A

The simplest idea is to assume that a priori we have some information on the marker variance. For instance, this can be  $\sigma_b^2$ . Thus, we may attach some importance to this value and use it as prior information for  $\sigma_{b_i}^2$ . A natural way of doing this is using an inverted chi-squared distribution with  $\nu_b$  degrees of freedom and scale parameter  $S_b^2 = \nu_b \sigma_b^2$

$$p(b_i|\sigma_{b_i}^2) = N(0, \sigma_{b_i}^2) \quad (8)$$

In the second stage, we postulate a prior distribution for the variance themselves:

$$p(\sigma_{b_i}^2|v_b, S_b^2) = S_b^2 \chi_{v_b}^{-1} \quad (9)$$

The value of  $\sigma_b^2$  should be set as

$\sigma_b^2 = \frac{v_b-2}{v_b} \frac{\sigma_g^2}{2 \sum_i p_i(1-p_i)}$  because the variance of a t distribution is  $\frac{v_b}{v_b-2}$ . It can be shown that this corresponds to a prior on the marker effects corresponding to a scaled t distribution (Gianola et al. 2009):

$$p(b_i|\sigma_b^2, v_b) = \sigma_b t(0, v_b) \quad (10)$$

which has the property of having “fat tails”. This means that large marker effects are more likely a priori compared to a normal distribution.

### Prior marker variance Bayes B

A very common thought was that there were not many causal loci. So a natural thinking is to consider that many markers do not have effect because they cannot trace causal loci. This originated the method known as BayesB, that simply states that the individual marker variance,  $\sigma_{b_i}^2$ , is potentially zero. Note that this cannot happen for BayesA as the a priori chi-squared distribution prevents any marker variance from being zero. This idea corresponds to a more complex prior as follows:

$$p(b_i|\sigma_{b_i}^2) = N(0, \sigma_{b_i}^2) \quad (11)$$

$$p(\sigma_{b_i}^2|S_b^2, v_b) = \begin{cases} 0 & \text{with probability } \pi, \\ S_b^2 \chi_{v_b}^{-1} & \text{with probability } 1 - \pi, \end{cases} \quad (12)$$

then when  $\sigma_{b_i}^2 = 0$  it follows that  $b_i = 0$ . BayesB has a complex algorithm because it does involve the computation of a complex likelihood and experience suggest that this methods is not very robust.

### Prior marker variance Bayes C

Habier et al. 2011 suggests the possibility of a simpler prior scheme where markers having an effect would be assigned a “common” variance such as  $\sigma_b^2$ . This is simpler to be explained by introducing additional variables  $\delta_i$  which explain if the i-th marker has an effect or not. In turn, these variables  $\delta$  have a prior distribution called Bernoulli with a probability  $\pi$  of being 1. Therefore the hierarchy of priors is:

$$p(\beta_j|\delta_i, \sigma_{b_i}^2, \pi) = \begin{cases} 0 & \text{with probability } \pi, \\ \sim N(0, \sigma_{b_i}^2) & \text{with probability } 1 - \pi, \end{cases} \quad (13)$$

$$p(\sigma_{b_i}^2 | v_b, S_b^2) = S_b^2 \chi_{v_b}^{-1} \quad (14)$$

where  $S_b^2 = \sigma_b^2 v_b$  with

$\sigma_b^2 = \frac{\sigma_g^2}{(1-\pi)2 \sum_i p_i(1-p_i)}$  because the variance of a t distribution is  $\frac{v_b}{v_b-2}$ . Experience shows that this prior hierarchy is more robust than BayesB (not shown), the reason being that, at the end (after fitting the data), the values of  $\sigma_{b_i}^2$  are little dependent on the prior. Thus the model may be correct even if the prior is wrong.

### Prior marker variance Bayesian Lasso

The Bayesian Lasso (Park and Casella 2008 ; Campos et al. 2009 ; A. Legarra, Robert-Granié, Croiseau, et al. 2011) suggests a different way to model the effect of markers. Instead of setting a priori some of them to 0, it sets them to very small values which corresponds to the following a priori distribution of markers:

$$p(b_i | \sigma_{b_i}^2) = N(0, \sigma_{b_i}^2) \quad (15)$$

In the second stage, we postulate a prior distribution for the variance themselves:

$$p(\sigma_{b_i}^2 | \lambda) = \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2} \frac{\sigma_{b_i}^2}{\sigma_e^2}\right) \quad (16)$$

In this case, a natural way of fitting the prior value of  $\lambda$  is as (Pérez et al. 2010)

$$\frac{2}{\lambda^2} = \frac{\sigma_g^2}{\sigma_e^2 2 \sum_i p_i(1-p_i)} \quad (17)$$

In this case,  $\lambda$  can be thought of as a ratio between marker variance and residual variance (signal-to-noise).

We can use any coding (centered, 101 or 012 or 210) in W for Bayesian methods. The estimated  $\hat{g}$  will be the same, the estimated  $\sigma_b^2$  or  $\pi$  will be the same, and the estimated genetic variance computed using, for instance,  $\sigma_g^2 = \sigma_b^2 2 \sum_i p_i(1-p_i)$ , will be the same too. These results are convenient because they assure us that any allele coding is convenient. However, this result does not apply to the all features. For instance, the standard deviation (and therefore, in animal breeding words, the “model-based” reliability) of estimated genetic values  $\hat{g}$  is not invariant to parameterization, because there will be a part of the overall mean absorbed, or not, by  $W\hat{b}$ . This implies that reports of the posterior variance of  $\hat{g}$  will depend on the allele coding.

### Prior marker variance Bayes R

The marker effects, b, are a priori assumed to be sampled from a mixture with a point mass at zero and univariate normal distributions conditional on common marker effect variance  $\sigma_\beta^2$ , and variance scaling factors,  $\gamma$ :

$$\beta_j | \pi, \sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2 \sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C \sigma_\beta^2) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases} \quad (18)$$

where  $\pi = (\pi_1, \pi_2, \dots, \pi_C)$  is a vector prior probabilities and  $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_C)$  is a vector of variance scaling factors for each of C marker variance classes. The  $\gamma$  coefficients are prespecified and constrain how the common marker effect variance  $\sigma_\beta^2$  scales within each mixture distribution. Typically  $\gamma = (0, 0.01, 0.1, 1.0)$ . and  $\pi = (0.95, 0.02, 0.02, 0.01)$ .

The prior distribution for the marker variance  $\sigma_\beta^2$  is assumed to be an inverse Chi-square prior distribution,  $\chi^{-1}(S_\beta, \nu_\beta)$ .

The proportion of markers in each mixture class  $\pi$  follows a Dirichlet  $(C, c + \alpha)$  distribution, where  $c$  is a vector of length C that contains the counts of the number of variants in each variance class and  $\alpha = (1, 1, 1, 1)'$ .

Using the concept of data augmentation, an indicator variable  $\mathbf{d} = (d_1, d_2, \dots, d_{m-1}, d_m)$ , is introduced, where  $d_j$  indicates whether the  $j$ 'th marker effect is zero or nonzero.

## Marker variance

The total genetic variance explained by markers can be expressed as (assuming linkage equilibrium):

$$\sigma_g^2 = 2 \sum_i p_i (1 - p_i) \hat{b}_i^2 \quad (19)$$

In most cases we do not know the marker effects, but may have some information prior variance then we can substitute the term  $\hat{b}_i^2$  by its expectation which is  $\sigma_{b_i}^2$  and therefore

$$\sigma_g^2 = 2 \sum_i p_i (1 - p_i) \sigma_{b_i}^2 \quad (20)$$

and if we assume that all markers have the same variance a priori then

$$\sigma_{b_i}^2 = \frac{\sigma_g^2}{2 \sum_i p_i (1 - p_i)} \quad (21)$$

which puts the prior variance of the markers as a function of the genetic variance of the population. This result is widely used in many applications in genomic prediction.

## 2.3 Estimation of model parameters

Bayesian linear regression methods use an iterative algorithm for estimating joint marker effects. Estimation of the joint marker effects depend on additional model parameters such as a probability of being causal ( $\pi$ ), an overall marker variance ( $\sigma_\beta^2$ ), and residual variance ( $\sigma_e^2$ ). Estimation of model parameters can be done using MCMC techniques by sampling from fully conditional posterior distributions or by using an EM based approximation (emBayesR) as a fast counterpart to a MCMC version of BayesR (Wang et al. 2015).

### 2.3.1 An MCMC algorithm for BayesR (mcBayesR)

An MCMC algorithm can be used to obtain estimates of parameters in BayesR models. The multiple linear regression model is parameterized in terms of  $\theta = (\sigma_e^2, \sigma_\beta^2, \gamma, \pi, b, d)$ . The full conditional sampling distributions for these parameters are presented below.

The joint posterior for all parameter in the multiple regression model can be written as :

$$f(\sigma_e^2, \sigma_\beta^2, \gamma, \pi, b, d|y) \propto (y|\sigma_e^2, \sigma_\beta^2, \gamma, \pi, b, d) f(b|d, \gamma, \sigma_\beta^2) f(d|\pi) f(\sigma_e^2) f(\sigma_\beta^2) f(\pi) \quad (22)$$

The parameters  $d_j$  and  $b_j$  are sampled jointly from their joint full conditional distributions, which can be written as the product of the full conditional distribution of  $b_j$  given  $d_j$  and the marginal full conditional distribution of  $d_j$ :

$$f_j(d_j, b_j|\theta_{-j}, \tilde{y}) = f_j(b_j|\theta_{-j}, d_j, \tilde{y}) f_j(d_j|\theta_{-j}, \tilde{y}) \quad (23)$$

Where  $\theta_{-j}$  is all parameters except  $d_j$  and  $b_j$ , and  $\tilde{y} = y - X_{-j}b_{-j}$  is the phenotype adjusted for all marker effects except for  $j$ 'th marker ( $b_j$ ).

The full conditional distribution for  $b_j$  can be written as:

$$f_j(b_j|\theta_{-j}, d_j, \tilde{y}) \propto N(C_j^{-1}r_j, C_j^{-1}) \quad (24)$$

where  $C_j^{-1} = \frac{\sigma_e^2}{x_j'x_j + \frac{\sigma_e^2}{\sigma_c^2}}$  and  $r_j = \frac{x_j'\tilde{y}}{\sigma_e^2}$

The marginal full conditional probability for the indicator variable of  $d_j$  is:

$$f_j(d_j = c|\theta_{-j}, \tilde{y}) = \frac{f_j(\tilde{y}|d_j = c, \theta_{-j}) f(d_j|\pi_c)}{\sum_{k=1}^C f_j(\tilde{y}|d_j = k, \theta_{-j}) f(d_j|\pi_k)} \quad (25)$$

where  $f_j(\tilde{y}|d_j = c, \theta_{-j}) = (\sigma_c^2)^{-0.5} (C_j^{-1})^{0.5} \exp[\frac{1}{2}C_j^{-1}r_j^2]$

From these expression probabilities for the categorical distribution for an arbitrary number of mixture components can be computed. Given these probabilities sample from a categorical distribution, which determines which class the variant will be sampled from. Conditional on the marker variance class sample the effect from the relevant normal distribution or give it a zero effect. To sample from the categorical distribution: 1) create a vector of cumulative probabilities calculated from above ordered by category, and 2) accept the lowest  $c$  such that the cumulative probability  $> u$ , where  $u$  is sample from a uniform distribution  $U(0, 1)$ .

The proportion of markers in each mixture class  $\pi$  follows a *Dirichlet*  $(C, c + \alpha)$  distribution, where  $\mathbf{c}$  is a vector of length  $C$  that contains the counts of the number of variants in each variance class and  $\alpha = (1, 1, \dots, 1)$ .

The full conditional posterior distribution for  $\sigma_\beta^2$  is a scaled inverse chi-squared distribution with  $v_\beta = \tilde{v}_\beta + q$

degrees of freedom and scale parameter  $S_\beta^2 = \frac{\tilde{v}_\beta \tilde{S}_\beta^2 + \sum_{j=1}^q \frac{b_j^2}{\gamma_{d_j}}}{v_\beta + q}$  where  $\tilde{v}_\beta$  and  $\tilde{S}_\beta^2$  are the prior degrees of freedom and scale parameters.

The full conditional posterior distribution for  $\sigma_e^2$  is a scaled inverse chi-squared distribution with  $v_e = \tilde{v}_e + n$  degrees of freedom and scale parameter  $S_e^2 = \frac{\tilde{v}_e \tilde{S}_e^2 + sse}{v_e + n}$  where  $\tilde{v}_e$  and  $\tilde{S}_e^2$  are the prior degrees of freedom and scale parameters and  $sse = y'y - b'r^* - b'X'y$ .

## 2.4 Mapping of variants

The value of  $1 - \pi$  (the number of markers having an effect) can be either fixed to a value or estimated from data. This is achieved in the last lines of the code above. How is this possible? Intuitively, we look at the number of markers who have ( $\delta_i = 1$ ) or not ( $\delta_i = 0$ ) an effect. Then we add a prior information on  $\pi$ . This comes in the form of a Beta( $a, b$ ) distribution, which is a distribution of fractions between 0 and 1, saying

that our fraction is a priori ‘like if’ we had drawn a black balls and b red balls from an urn to make  $\pi = \frac{a}{a+b}$ . The total genetic variance explained by markers in the Bayes C model is equal to

$$\sigma_g^2 = \sigma_b^2(1 - \pi)2 \sum_i p_i(1 - p_i) \quad (26)$$

Thus, the same total genetic variance can be achieved with large values of  $\sigma_b^2$  and small values of  $(1 - \pi)$  or the opposite. This implies that there is a confusion between both, and it is not easy to find out how many markers should be in the model. How can we declare significance? There is no such thing as p-values. We may though use the Bayes Factor (Wakefield 2009 ; Varona 2010) :

$$BF_i = \frac{1 - \pi}{\pi} \frac{p(\delta_i = 1|y)}{1 - p(\delta_i = 1|y)} \quad (27)$$

What thresholds should we use for BF? We can use a scale adapted by (Kass and Raftery 1995) sometimes used in QTL detection (Varona, García-Cortés, and Pérez-Enciso 2001 ; Vidal et al. 2005): BF= 3-20 ‘suggestive’ BF= 20-150 ‘strong’ BF>150 ‘very strong’

Something remarkable is that there is no need for multiple testing (Bonferroni) correction because all SNP were introduced at the same time, and the prior already penalizes their estimates (Wakefield 2009).

## 2.5 Extensions to Multiple Traits and Multiple Components

Bayesian linear regression models have been extended to multiple trait analyses. A multiple trait analysis is a natural choice in quests for understanding and dissecting genetic correlations between traits using genetic markers, e.g., evaluating whether pleiotropy or linkage disequilibrium are at the root of between-trait associations. Furthermore pleiotropy among traits can be utilised to increase accuracy of genomic predictions.

Multiple trait Bayesian linear regression models can implemented under an restrictive assumption that a locus simultaneously affects all the traits or none of them. However this assumption is not biologically meaningful, especially in multiple trait analyses involving many traits. This assumption of genetic architecture is violated if some loci have no effect on at least one of the traits while having an effect on the remaining traits.

Therefore a general multiple trait Bayesian regression model allowing a broad range of mixture priors have been proposed. In this model a locus is allowed to affect any combination of traits, e.g., in a 2-trait analysis, the “restrictive” model only allows two situations, whereas in the general model allow all  $2^2 = 4$  situations. This model is particularly interesting because it provides insight into whether markers affect all, some, or none of the traits addressed. For example, the proportion of markers in each of the (0,0), (1,0), (0,1) and (1,1) categories, where (0,0) means “no effect,” and (1,1) denotes “effect” on both traits.

A further extensions Bayesian regression model is to partitioning the markers into multiple sets based on prior knowledge on genomic features.

### 2.5.1 Prior distributions for multiple traits

In practice, genetic variances from previous conventional analyses are usually used to construct priors for marker effect variances. For single trait analyses, under some assumptions, it can be shown that the marker effect variance  $XX$  can be obtained as  $XX(2)$  where  $XX$  is the genetic variance,  $XX$  is the allele frequency for locus  $j$ , and  $\pi$  is the probability that a marker has a null effect (Habier et al. 2007; Gianola et al. 2009; Fernando and Garrick 2013). Following a similar strategy, the marker effect covariance matrix  $XX$  in a two-trait analysis can be obtained as  $XX(3)$  where  $XX$  is the genetic covariance matrix and  $XX$   $XX$  and



XX are the probabilities a marker has null effects on the first trait but not the second trait, on the second trait but not the first trait, or on neither trait. Thus the probability that a marker has an effect on the first trait can be obtained as  $\pi$  which is the denominator of the upper left element in (3). This strategy relating marker effect covariance matrix to genetic covariance matrix can be readily extended to  $>2$  traits. Note that positive definite matrix  $\Sigma$  may result in negative definite matrix  $\Sigma$  using (3), especially when the prior for the probability a marker has null effects is far from the real value. In that case, the diagonal elements of  $\Sigma$  which are the marker effect variances for different traits, can be obtained using (2), where  $\pi$  may be estimated from previous single-trait analyses, and the off-diagonal elements of  $\Sigma$  may be set to zero to guarantee positive definiteness of  $\Sigma$ .

### 2.5.2 Multi-trait variable selection

In regard to a single trait, a locus either has an effect, or it does not. Hence, the scalar parameter  $\pi$  (and its complement  $\Sigma$ ) completely defines this circumstance. In a multi-trait setting, it is conceivable that loci that influence one trait, may or may not influence other traits. In that circumstance, a vector  $\Sigma$  is required to define the genetic architecture. The number of parameters that constitute the vector  $\Sigma$  is  $\Sigma$  which grows rapidly with the number of traits. In most cases, the researcher will have little or no knowledge of the likely extent of pleiotropy of loci that influence two traits, other than knowing or having an estimate of the genetic covariance. There are two simple ways to reduce this complexity in priors.

First, one can assume, as did Jia and Jannink (2012), that, in the context of variable selection, a locus should be selected for all of the traits or selected for none of the traits, reducing the required probabilities to being analogous to the single trait  $\pi$  and  $\Sigma$ . This approach has the advantage of simplicity, but the disadvantage that many effects might need to be estimated for loci that have no effect on a trait, and this may erode the accuracy of prediction. This should not be a problem for asymptotically large datasets, as in that case the fitted locus effects should converge to zero for those traits not influenced by that locus.

A second simple way to accommodate the multiple trait circumstance is to assume the  $\Sigma$  parameters can be derived from  $t$  trait-specific parameters. However, when the probability that a single trait locus has an effect is small for each of two or more traits, the pair-wise probability that a locus affects all the traits will be the product of those small probabilities, making it very difficult for loci to enter the model for all traits simultaneously.

The better way to solve this problem is to use a hyper-parameter  $\Sigma$  that completely defines the alternative models that are required to capture all the alternative forms of genetic architecture. We have shown here how this can be done, with two alternative Gibbs sampling strategies. One involves single-site sampling for one locus and trait at a time. The other samples all the alternative combinations of effects for one locus considering all traits simultaneously. We have shown that both are practical with real data and can result in improved accuracies of prediction in certain circumstances in terms of genetic architecture and size of dataset.

### 2.5.3 Posterior sampling distributions for multiple traits

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \left( \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix} + I \otimes G^{-1}E \right)^{-1} \begin{bmatrix} X_1'y_1 \\ X_2'y_2 \end{bmatrix} \quad (28)$$

$$G = \begin{bmatrix} \sigma_{s_1}^2 & \sigma_{s_{12}}^2 \\ \sigma_{s_{21}}^2 & \sigma_{s_2}^2 \end{bmatrix} \quad (29)$$

$$E = \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}}^2 \\ \sigma_{e_{21}}^2 & \sigma_{e_2}^2 \end{bmatrix} \quad (30)$$

$$\sigma_{e_{12}}^2 = \sigma_{e_{21}}^2 = 0$$

The covariance matrix  $\mathbf{G}$  for the marker effect is a priori assumed to follow an inverse Wishart distribution  $IW(\tilde{S}_\beta^2, \tilde{\nu}_\beta)$  where  $\tilde{\nu}_\beta$  and  $\tilde{S}_\beta^2$  (t by t matrix) are the prior degrees of freedom and scale parameters.

The full conditional distribution for  $\mathbf{G}$  the covariance matrix for the jth marker is also an inverse Wishart distribution of the form:

$$IW(\tilde{S}_\beta^2 + \mathbf{b}\mathbf{b}', \tilde{\nu}_\beta + m), \text{ where } \mathbf{b} \text{ is a matrix (m x t) of sampled marker effects}$$

or

$$IW(\tilde{S}_\beta^2 + b_j b_j', \tilde{\nu}_\beta + 1), \text{ where } b_j \text{ is a vector (t x 1) of sampled marker effects.}$$

Difference between the two sampling distributions is that the former is the same for all markers whereas the latter is more flexible as it allows the markers to their own covariance matrix.

The covariance matrix  $\mathbf{E}$  for the residual effects is a priori assumed to follow an inverse Wishart distribution  $IW(\tilde{S}_e^2, \tilde{\nu}_e)$  where  $\tilde{\nu}_e$  and  $\tilde{S}_e^2$  (t by t matrix) are the prior degrees of freedom and scale parameters.

The full conditional distributions for the covariance matrix  $\mathbf{E}$  for residuals is also an inverse Wishart distribution of the form:

$$IW(\tilde{S}_e^2 + \mathbf{e}\mathbf{e}', \tilde{\nu}_e + m), \text{ where } \mathbf{e} \text{ is a matrix (n x t) of residual effects}$$

The full conditional distribution for  $b_j$  can be written as:

$$f_j(b_j | \theta_{-j}, d_j, \tilde{y}) \propto N(C_j^{-1} r_j, C_j^{-1}) \quad (31)$$

where  $C = D_j' E^{-1} D_j x_j' x_j + G^{-1}$  and  $r_j = x_j' \tilde{y} E^{-1} D_j = \tilde{b}_j x_j' x_j E^{-1} D_j$

where  $D_j$  is a diagonal matrix whose kth diagonal entry is an indicator variable indicating whether the marker jth effect for trait k is zero or nonzero.

The marginal full conditional probability for the indicator variable of  $d_j$  is:

$$f_j(d_j = c | \theta_{-j}, \tilde{y}) = \frac{f_j(\tilde{y} | d_j = c, \theta_{-j}) f(d_j | \pi_c)}{\sum_{k=1}^C f_j(\tilde{y} | d_j = k, \theta_{-j}) f(d_j | \pi_k)} \quad (32)$$

where  $f_j(\tilde{y} | d_j = c, \theta_{-j}) = |C_j^{-1}|^{0.5} \exp\left[-\frac{1}{2} r_j' C_j^{-1} r_j\right]$

In the most general case, any marker effect might be zero for any possible combination of t traits resulting in  $2^t$  possible combinations of  $d_j$ : For example, in a t=2 trait model, there are  $2^2 = 4$  combinations for  $d_j$ : (0,0), (0,1), (1,0), (1,1). Suppose, in general, we use numerical labels 1,2,...,l for the  $2^t$  possible outcomes for  $d_j$  then the prior for  $d_j$  is a categorical distribution.

## 2.6 Extensions to Summary Statistics

The multiple regression model described above can be related to the estimates of the regression coefficients from m simple linear regressions  $b_{marg}$  from a standard GWAS, by multiplying by  $D^{-1} X'$  where  $D = \text{diag}(x_1' x_1, x_2' x_2, \dots, x_m' x_m)$  such that:

$$D^{-1} X' y = D^{-1} X' X b + D^{-1} X' e \quad (33)$$

Noting that  $b_{\text{marg}} = D^{-1}X'y$  is a vector ( $m \times 1$ ) of least-squares marginal regression effect estimates and the LD correlation matrix between all genetic markers  $B = D^{-0.5}X'XD^{-0.5}$ . The key parameter of interest in the multiple regression model are the marker effects. These can be obtained by solving an equation system similar to:

$$\hat{b} = \left( X'X + I \frac{\sigma_e^2}{\sigma_\beta^2} \right)^{-1} X'y \quad (34)$$

$$\hat{b} = (X'X + I\lambda)^{-1} X'y \quad (35)$$

$$\lambda = \left( \frac{\sigma_e^2}{\sigma_{b_1}^2}, \frac{\sigma_e^2}{\sigma_{b_2}^2}, \dots, \frac{\sigma_e^2}{\sigma_{b_m}^2} \right) \quad (36)$$

$\sigma_{b_j}^2$  can take values  $\gamma_c \sigma_\beta^2$  where  $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$  depending on the state of the indicator variable  $d$ .

In order to solve this equation system individual level data (genotypes  $X$  and phenotypes  $y$ ) is required. If these are not available, it is possible to reconstruct  $X'y$  and  $X'X$  from summary statistics and LD reference panel.  $X'X$  is derived from an LD matrix  $\mathbf{B}$  (from population matched reference) and summary statistics:

$$X'X = D^{0.5}BD^{0.5} \quad (37)$$

where  $D_i = \frac{1}{\hat{\sigma}_{b_i}^2 + \hat{b}_i^2/n_i}$  if the genotypes have been centered to mean 0  $D_i = n_i$  if the genotypes have been centered to mean 0 and scaled to unit variance.  $X'y$  is derived from univariate marker effects ( $b_{\text{marg}}$ ):

$$b_{\text{marg}} = D^{-1}X'yX'y = Db_{\text{marg}} \quad (38)$$

## Reference LD matrix construction

The summary statistics methods used require the construction of a reference LD correlation matrix. Typically this is done through the use of a fixed 1–10-Mb window approach, as in GCTA-SBLUP or LDpred, which sets LD correlation values outside this window to zero. Zhu and Stephens detail the reasons for using the shrinkage estimator of the LD matrix, which shrinks the off-diagonal entries of the LD correlation matrix towards zero and is required for the implementation based on summary statistics. Experimentation with different types of sparse LD correlation matrices led to the conclusion that the shrinkage estimator was the most stable for SBayesR implementation. Briefly, each element of the reference LD correlation matrix  $B_{ij}$  is shrunk by the factor  $\exp(-\varrho_{ij}/2m)$ , where  $m$  is taken to be the sample size used to generate the genetic map, ( $\varrho_{ij}$ ) is an estimate of the population-scaled recombination rate between SNPs  $i$  and  $j$  taken as  $\varrho_{ij} = 4N_e c_{ij}$ , for  $N_e$  the effective population size and  $c_{ij}$  the genetic distance between sites  $i$  and  $j$  in centimorgans as stated in Li and Stephens. LD matrix entries are set to zero if  $\exp(-\varrho_{ij}/2m)$  is less than a user-chosen cutoff.

Genetic distance between sites is derived from the genetic map files containing interpolated map positions for the CEU population generated from the 1000G OMNI arrays (Data availability). The calculation of the shrunk LD matrix requires the effective population sample size, which we set to be 11,400 (as in Zhu and Stephens), the sample size of the genetic map reference, which corresponds to the 183 individuals from

the CEU cohort of the 1000G and the hard threshold on the shrinkage value, which we set to  $10^{-3}$ . This threshold gave a good balance between computational efficiency and accuracy with, on average, each SNP having a window width of 10.6 Mb (SD=5.6Mb) across the autosomes. The shrunk LD matrix is stored in a sparse matrix format (ignoring matrix elements equal to 0) for efficient SBayesR computation. Currently, the LD matrix construction can only be performed with PLINK hard-call genotypes.

### 2.6.1 McMC algorithm for a single trait Bayesian model based on summary statistics

The following steps are required, 1) reconstruct  $X'y$  and  $X'X$  from summary statistics and LD reference panel as explained above, 2) initialise parameters, 3) sampling of model parameters from full conditional posterior distributions described above.

Initialize parameters:

$$b = 0$$

$$\sigma_e^2 = \frac{sse}{n - 1}$$

$$sse = y'y - 2b'X'y + b'X'Xb = y'y \quad (\text{given that } b = 0 \text{ initially})$$

$$y'y = \frac{1}{m} \sum_{j=1}^m (y'y)_j$$

$$(y'y)_j = \hat{\sigma}_{b_i}^2 x'_j x_j (n - 2) + \hat{b}_i^2 x'_j x_j$$

$$h_{snp}^2 = 0.5$$

$$\sigma_g^2 = \sigma_e^2 h_{snp}^2$$

$$\sigma_\beta^2 = \frac{\sigma_g^2}{(1 - \pi_{c=1}) \sum_{j=1}^m 2p_j (1 - p_j)}$$

$$\pi = (0.995, 0.002, 0.002, 0.001)$$

$$\gamma = (0, 0.01, 0.1, 1.0)$$

$$\tilde{v}_\beta = \tilde{v}_e = 4$$

$$\tilde{S}_\beta^2 = \frac{(\tilde{v}_\beta - 2) \sigma_\beta^2}{\tilde{v}_\beta} = \frac{\sigma_\beta^2}{2}$$

$$\tilde{S}_e^2 = \frac{(\tilde{v}_e - 2) \sigma_e^2}{\tilde{v}_e} = \frac{\sigma_e^2}{2}$$

$$r^* = X'y - X'Xb$$

$$\left( r^* = X'y \text{ because initially } b = 0 \right)$$

Gibbs sampler:

for i = 1 to number of iterations do

for j = 1 to number of markers do

$$r_j = r_j^* + x_j'x_j b_j^i \quad (\text{rhsformmarkerj})$$

$$\sigma_c^2 = \sigma_\beta^2 \gamma_{d_j=c} \text{ for each of C classes } (e.g. 4 \text{ classes } \gamma = (0, 0.01, 0.1, 1.0))$$

$$l_{jc} = x_j'x_j + \frac{\sigma_e^2}{\sigma_c^2} \quad (\text{lhsformmarkerjinclassc})$$

$$\log L_c = 0.5 \left[ \log \left( \frac{\sigma_e^2}{\sigma_c^2 l_{jc}} \right) + \frac{\sigma_e^2}{\sigma_c^2} \left( \frac{r_j^2}{\sigma_c^2 l_{jc}} \right) \right] + \log(\pi_c) \quad (\text{loglikelihoodformarkerjinclassc})$$

$$P(d_j = c | \theta_{-j}, \tilde{y}) = \frac{1}{\sum_{k=1}^C \exp[\log L_k - \log L_c]} \quad (\text{posteriorprobabilityformarkerjinclassc}) \quad \text{Sample variance class membership, } d_j,$$

Sample marker effect,  $b_j^{i+1}$ , from  $N\left(\frac{r_j}{l_{jc}}, \frac{\sigma_e^2}{l_{jc}}\right)$  based on variance class membership  $d_j$

$$\text{Update right hand side } r_j^{*(i+1)} = r_j^{*(i)} - X_j'x_j (b_j^{i+1} - b_j^i) \text{ where } X_j'x_j \text{ is the } j\text{th column of } X'X$$

enddo

Sample  $\sigma_\beta^2$  from scaled inverse chi-squared distribution with  $v_\beta = \tilde{v}_\beta + q$  degrees of freedom and scale

$$\text{parameter } S_\beta^2 = \frac{\tilde{v}_\beta \tilde{S}_\beta^2 + \sum_{j=1}^q \frac{b_j^2}{\gamma_{d_j}}}{\tilde{v}_\beta + q} \text{ where } \tilde{v}_\beta \text{ and } \tilde{S}_\beta^2 \text{ are the prior degrees of freedom and scale parameters.}$$

Sample  $\sigma_e^2$  from scaled inverse chi-squared distribution with  $v_e = \tilde{v}_e + n$  degrees of freedom and scale

$$\text{parameter } S_e^2 = \frac{\tilde{v}_e \tilde{S}_e^2 + sse}{\tilde{v}_e + n} \text{ where } \tilde{v}_e \text{ and } \tilde{S}_e^2 \text{ are the prior degrees of freedom and scale parameters and } sse = y'y - b'r^* - b'X'y.$$

Sample  $\pi$  from a *Dirichlet*  $(C, c + \alpha)$ , where  $c$  is a vector of length C that contains the counts of the number of variants in each variance class and  $\alpha = (1, 1, \dots, 1)$ .

$$\text{Calculate } \sigma_g^2 = \frac{mss}{n} \text{ where } mss = b'X'y - b'r$$

$$\text{Calculate } h_{snp}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

enddo

### 3 Bayesion Linear Regression Model for Individual Effects

#### 3.1 Statistical model and model parameters

We begin with a description of the general two-trait model since the simpler models are special cases. The general two-trait model partitions the total phenotypic variance into a component explained by genetic marker effects,  $g$ , and a component due to residual effects that are not associated with markers,  $e$ . Additionally, the component due to genetic markers  $g$  is subdivided into contributions from  $C$  marker sets  $g_i$ ,  $i = 1, \dots, C$ .

The model for trait 1 is assumed to be

$$y_1 = \mu_1 + \sum_{i=1}^C g_{1i} + e_1, \quad (39)$$

and similar for trait 2,

$$y_2 = \mu_2 + \sum_{i=1}^C g_{2i} + e_2, \quad (40)$$

In these expressions, the  $\mu$ 's are scalar means for each trait, the  $g$ 's are contributions to genetic effects from each of  $C$  marker sets that can be associated with genetic marker information (genomic effects or genomic values), the  $e$ 's represent a residual component that cannot be captured by regression on markers.

The genomic value for marker set  $k$  is defined as the sum of the effects of all the markers in marker set  $k$ ,

$$\begin{aligned} g_{1k} &= \sum_{i=1}^{p_k} w_{ijk} b_{1ik} \\ &= w'_{jk} b_{1k}, \end{aligned} \quad (41)$$

with an equivalent expression for trait 2. In (41),  $p_k$  is the number of markers in marker set  $k$ , the scalar  $w_{ijk}$  is the observed (centered and scaled) label for marker  $i$  in individual  $j$  for marker set  $k$ , and  $b_{1ik}$  is the effect of marker  $i$  for marker set  $k$  of trait 1.

The  $n$  by 1 vectors of genomic effects for trait 1 and trait 2 for marker set  $k$  are  $g_{1k} = W_k b_{1k}$  and  $g_{2k} = W_k b_{2k}$ , respectively, where  $W_k = \{w_{ijk}\}$  is the observed  $n$  by  $p_k$  matrix of marker genotypes of marker set  $k$ . The joint distribution of vectors  $b_{1k}$  and  $b_{2k}$  is assumed to be

$$\begin{bmatrix} b_{1k} \\ b_{2k} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I\sigma_{b_{1k}}^2 & I\sigma_{b_{1k}b_{2k}} \\ I\sigma_{b_{1k}b_{2k}} & I\sigma_{b_{2k}}^2 \end{bmatrix} \right) \quad (42)$$

where the  $I$ 's represent  $p_k$  by  $p_k$  identity matrices,  $\sigma_{b_{1k}}^2$  ( $\sigma_{b_{2k}}^2$ ) is the prior variance of marker effects for trait 1 (trait 2) in marker set  $k$ , and  $\sigma_{b_{1k}b_{2k}}$  is their prior covariance.

Due to the centering the rank of  $W_k$  is  $n - 1$ . It follows from these assumptions and from standard properties of the multivariate normal distribution that the model for the joint distribution of genomic values in trait 1 and 2 is the singular multinormal ( $SN$ ) distribution,

$$\begin{bmatrix} g_{1k} \\ g_{2k} \end{bmatrix} \sim SN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_k \sigma_{g_{1k}}^2 & G_k \sigma_{g_{1k}g_{2k}} \\ G_k \sigma_{g_{1k}g_{2k}} & G_k \sigma_{g_{2k}}^2 \end{bmatrix} \right), \quad k = 1, \dots, C. \quad (43)$$

Above,  $\sigma_{g_{1k}}^2$  is the genomic variance in trait 1 for marker set  $k$ ,  $\sigma_{g_{2k}}^2$  is the genomic variance in trait 2 for marker set  $k$ ,  $\sigma_{g_{1k}g_{2k}}$  is the genomic covariance between trait 1 and 2 for marker set  $k$ ,  $G_k = \frac{1}{p_k} W_k W_k'$  is the genomic relationship matrix of rank  $n - 1$  for marker set  $k$ , and  $SN$  denotes the singular normal distribution (details in the Supplementary Methods). A more compact notation for (43) is

$$g_k = (g'_{1k}, g'_{2k})' \sim SN(0, G_k \otimes V g_k) \quad (44)$$

where

$$V g_k = \begin{bmatrix} \sigma_{g_{1k}}^2 & \sigma_{g_{1k}g_{2k}} \\ \sigma_{g_{1k}g_{2k}} & \sigma_{g_{2k}}^2 \end{bmatrix}. \quad (45)$$

Similarly, collecting the  $e$ 's in vectors with  $n$  elements for each trait joint distribution is assumed to be

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I\sigma_{e_1}^2 & I\sigma_{e_1 e_2} \\ I\sigma_{e_1 e_2} & I\sigma_{e_2}^2 \end{bmatrix} \right), \quad (46)$$

The identity matrix  $I$  is of order  $n$  by  $n$  in (46). The phenotypes for trait 1 and 2,  $y_1$  and  $y_2$ , each with  $n$  elements, are conditionally normally distributed, given  $g$ , and  $e$ , and take the form

$$y_1 | \mu_1, g_1 \sim N \left( 1\mu_1 + \sum_{i=1}^C g_{1i}, I\sigma_{e_1}^2 \right), \quad (47)$$

$$y_2 | \mu_2, g_2 \sim N \left( 1\mu_2 + \sum_{i=1}^C g_{2i}, I\sigma_{e_2}^2 \right), \quad (48)$$

where  $\sigma_{e_1}^2$  ( $\sigma_{e_2}^2$ ) is the residual variance for the trait 1 (trait 2).

### 3.2 Extensions to multiple components

The model specified quantify the relative contribution from each of the marker sets by fitting separate variance components for each of those marker sets. Here we describe the basis for this partitioning. The point of departure is to assume that the  $2n_t$  by 1 vector of genomic effects  $g = (g'_1, g'_2)'$  is equal to the sum of the vectors of genomic effects belonging to  $C$  marker set components. Then

$$g = \sum_{i=1}^C g_i, \text{Var}(g|W) = \sum_{i=1}^C \text{Var}(g_i|W_i) = \sum_{i=1}^C \frac{1}{p_i} W_i W_i' \otimes V_{g_i} \quad (49)$$

where  $g_i$  is the  $2n$  by 1 vector of genomic values for trait 1 and 2, associated with component  $i$ ,  $i = 1, \dots, C$ ,  $p_i$  is the number of marker genotypes in component  $i$ , and  $V_{g_i}$  is the 2 by 2 genomic covariance matrix for marker set  $i$ . This holds under the model assumption that the components of the vector of marker effects  $b = (b'_1, \dots, b'_i, \dots, b'_C)'$  are realisations from  $b_i \sim N(0, I \otimes V_{b_i})$ , with  $\text{Cov}(b_i, b'_j) = 0$ ,  $i, j = 1, \dots, C$ ;  $i \neq j$ . Then  $\text{Cov}(g_i, g'_j|W) = \text{Cov}(W_i b_i, b'_j W'_j) = 0$ . In these expressions,  $b_i = (b'_{1i}, b'_{2i})'$  is the vector of marker effects of marker set component  $i$ , and

$$V_{b_i} = \begin{bmatrix} \sigma_{b_{1i}}^2 & \sigma_{b_{1i}b_{2i}} \\ \sigma_{b_{1i}b_{2i}} & \sigma_{b_{2i}}^2 \end{bmatrix}. \quad (50)$$

Two simple two-trait models are implemented that partition the total genomic variance into multiple component defined by marker sets. Details of the prior distributions of the parameters of the Bayesian models and of the Markov chain Monte Carlo algorithm can be found in the Supplementary Methods.

#### 3.2.1 Inferences

For single trait analysis, inferences are reported in terms of ratios for the  $i$ th marker sets defined as:

$$h_{g_i}^2 = \frac{\sigma_{g_i}^2}{\sigma_g^2}, \quad i = 1, \dots, C. \quad (51)$$

with  $\sigma_g^2 = \text{Var}(\sum_{i=1}^C g_i)$ . When the elements of  $W$  are scaled so that the average of the diagonal elements of  $WW'$  is equal to 1, this ratio quantifies the proportion of total genomic variance, captured by genetic

marker information associated with marker set  $i$ . The ratio involving marker effects from all the marker sets for trait 1 is

$$h_{g_1}^2 = \frac{\sigma_{g_1}^2}{\sigma_{g_1}^2 + \sigma_{e_1}^2} \quad (52)$$

with a similar expression for trait 2.

We also report within marker set genomic correlations

$$r_{g_i} = \frac{\sigma_{g_{i_1}g_{i_2}}}{\sigma_{g_{i_1}}\sigma_{g_{i_2}}}, \quad i = 1, \dots, C, \quad (53)$$

and the total genomic correlation between traits, defined as

$$r_{g_i} = \frac{Cov\left(\sum_{i=1}^C g_{1_i}, \sum_{i=1}^C g_{2_i}\right)}{\sqrt{\sigma_{g_1}^2 \sigma_{g_2}^2}} \quad (54)$$

The absence of symmetry of posterior distributions is well captured by the Bayesian McMC methods. In a situation like this, summarizing results in the form of posterior means and standard deviations would be misleading. In the presence of asymmetry, posterior means are poor indicators of points of high probability mass. We therefore summarize inferences in terms of posterior modes and posterior intervals.

### 3.3 Estimation of model parameters

The models are implemented using empirical Bayesian methods. The models were implemented using a Bayesian, Markov chain Monte Carlo approach, whereby the hyperparameters of the dispersion parameters of the Bayesian model were estimated by maximum likelihood, and conditional on these, the model was fitted using Markov chain Monte Carlo. To illustrate, consider determining a value of the scale parameter of an inverse chi square distribution, which is the prior assumed for the variance components. This is achieved by equating the mode of the inverse chi square prior density, for a fixed value of the degrees of freedom, to the maximum likelihood estimate and solving for the scale parameter. The decision to use this approach rather than classical likelihood is based on the need to obtain measures of uncertainty that account as much as possible for the limited amount of information in the data without resorting to large sample theory, which is inherent in classical likelihood. This is relevant if the data are of limited size and results in marginal posterior distributions that are expected to be asymmetric, particularly in the case of the more complex, general two-trait model.

A number of technical details are presented in this section. These include a formal derivation of the probability model for the variance between traits and details of a spectral decomposition that plays an important computational role in the Markov chain Monte Carlo strategy implemented, as well as a full description of the Bayesian models.

#### 3.3.1 The spectral decomposition

The Bayesian implementation of the model is facilitated making use of the following spectral decomposition performed within each chromosome segment (we drop subscripts that refer to marker sets to avoid clotting the notation). Let

$$g_1 = Wb_1g_2 = Wb_2 \quad (55)$$



where  $W$  is of order  $n \times p$  and  $g_m$  is of order  $n \times 1$ . The spectral decomposition of  $WW'$  (for each marker set) is

$$WW' = U\Delta U' = \sum_{i=1}^{n_t} \lambda_i U_i U_i', \quad (56)$$

where  $U = [U_1, U_2, \dots, U_{n_t}]$ , of order  $n \times n$  is the matrix of eigenvectors of  $WW'$ ,  $U_j$  is the  $j$ th column (dimension  $n \times 1$ ), and  $\Delta$  is a diagonal matrix with elements equal to the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_{n_t}$  associated to the  $n_t$  eigenvectors. Since  $WW'$  is positive semidefinite the eigenvalues are  $\lambda_i \geq 0$ ,  $i = 1, 2, \dots, n$ . The eigen vectors satisfy  $U'U = UU' = I$ .

Define the  $n \times n$  matrix  $G = \frac{1}{p}WW' = \frac{1}{p}WW'$  and write this as

$$\begin{aligned} G &= \frac{1}{p}WW' \\ &= \frac{1}{p}U\Delta U' \\ &= UDU' \end{aligned}$$

where

$$D = \frac{1}{p}\Delta.$$

Matrix  $G$  (peculiar to each marker set) is singular and the diagonal matrix  $D$  (also peculiar to each marker set) contains only  $n - 1$  positive eigenvalues.

### 3.3.2 The case of singular $G$

Due to the singularity of  $G$ ,  $[g|W, \sigma_g^2]$  does not follow a multivariate normal distribution but a singular normal distribution instead. The singular normal density is

$$p(g_1|W, \sigma_{g1}^2) = \frac{1}{(2\pi)^{\frac{n-1}{2}} (\lambda_1 \sigma_{g1}^2 \dots \lambda_{n-1} \sigma_{g1}^2)^{\frac{1}{2}}} \exp\left(-\frac{g_1' G^- g_1}{2\sigma_{g1}^2}\right) \quad (57)$$

where the  $n - 1$   $\lambda$ 's are the non-zero eigenvalues of  $G$  and  $G^-$  is any generalised inverse of  $G$ . One choice of generalised inverse of  $G$  is

$$G^- = U D^- U' \quad (58)$$

where  $U$  and  $D$  are of order  $n_t$  by  $n_t$

$$D^- = \frac{1}{p} \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \dots & \dots & 0 \\ 0 & \ddots & \dots & \dots & 0 \\ \vdots & \vdots & \frac{1}{\lambda_{n-1}} & \dots & 0 \\ \vdots & \vdots & \dots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \frac{1}{p} \begin{bmatrix} D_1^{-1} & 0 \\ 0' & 0 \end{bmatrix}$$

Above,  $D_1 = \text{diag}(\lambda_i)_{i=1}^{n-1}$ , a diagonal matrix of dimension  $n-1$  by  $n-1$  that contains the nonzero eigenvalues  $\lambda_i$ . The remaining elements of  $D^-$  are all equal to zero.

### 3.3.3 A probabilistically equivalent reparameterisation of the random regression model

We describe a reparameterisation of the original two-trait model that simplifies the Markov chain Monte Carlo computations.

For each marker set (omitting the subscripts) define the row vector random variable  $\alpha' = (\alpha'_1, \alpha'_2)$  of dimension 1 by  $2n$ , with vectors  $\alpha_1$  and  $\alpha_2$ , each of dimension  $n$  by 1 associated with trait 1 and 2, with distribution

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim SN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D\sigma_{g_1}^2 & D\sigma_{g_1 g_2} \\ D\sigma_{g_1 g_2} & D\sigma_{g_2}^2 \end{bmatrix} \right) \quad (59)$$

where  $D$  is a diagonal matrix (associated with a particular marker set) of dimension  $n \times n$ , with has eigenvalues  $\lambda_i$ ,  $i = 1, \dots, n$  as diagonal elements, of which the first  $n - 1$  are positive and the rest are equal to zero. We define for each marker set

$$V_g = \begin{bmatrix} \sigma_{g_1}^2 & \sigma_{g_1 g_2} \\ \sigma_{g_1 g_2} & \sigma_{g_2}^2 \end{bmatrix}. \quad (60)$$

In a particular marker set, the random variables

$$\begin{pmatrix} U\alpha_1 \\ U\alpha_2 \end{pmatrix} \sim SN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} UDU'\sigma_{g_1}^2 & UDU'\sigma_{g_1 g_2} \\ UDU'\sigma_{g_1 g_2} & UDU'\sigma_{g_2}^2 \end{bmatrix} \right) \quad (61)$$

and

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \sim SN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{p}WW'\sigma_{g_1}^2 & \frac{1}{p}WW'\sigma_{g_1 g_2} \\ \frac{1}{p}WW'\sigma_{g_1 g_2} & \frac{1}{p}WW'\sigma_{g_2}^2 \end{bmatrix} \right) \quad (62)$$

have the same distribution since  $UDU' = \frac{1}{p}WW'$ . Here  $p$  is the number of genetic markers in the marker set. Therefore the structure can be written as

$$y_1 | \mu_1, \sum_{i=1}^C \alpha_{1i} \sim N \left( 1\mu_1 + \sum_{i=1}^C U_i \alpha_{1i}, I\sigma_{1e}^2 \right), \quad (63)$$

$$y_2 | \mu_2, \sum_{i=1}^C \alpha_{2i} \sim N \left( 1\mu_2 + \sum_{i=1}^C U_i \alpha_{2i}, I\sigma_{2e}^2 \right) \quad i = 1, \dots, C. \quad (64)$$

### 3.3.4 Prior distributions

The prior distribution of the  $\mu$ 's is  $N(0, 10^5)$ . The prior distributions of  $V_g$  and  $V_h$  are scaled inverted Wishart with degrees of freedom set equal to 2.5 and scale parameters  $P_g$  and  $P_h$ , respectively. The degrees of freedom generate a proper distribution with overdispersed values. The prior distributions of  $\sigma_{r_m}^2$  and  $\sigma_{r_f}^2$  are scale inverted chi square densities. The degrees of freedom of these densities are set equal to 1.0 which leads to vague prior information. For example, when the scale is equal to 0.1, the modal value of the prior distribution is 0.03 and the prior probability that the variance component is smaller than this value is 56. The prior probability that the variance component is between 0.03 and 0.3 is 29.

The scale parameters of all these distributions are estimated using maximum likelihood. This involved obtaining maximum likelihood estimates of the two-trait model (for Models  $H$  and  $G$ , each sex considered as one trait), and then equating these estimates to the mode of the relevant prior distribution, written as a function of the scale parameter. In the case of the general two-trait model, single-trait likelihoods were fitted instead. The off-diagonal elements of the scale parameter of inverse Wishart distributions were set equal to zero.

### 3.3.5 Fully conditional posterior distributions

The fully conditional posterior distributions of a parameter  $\theta$  is denoted  $[\theta|All, z]$  where  $All$  denotes all the parameters of the model except  $\theta$ . Here we sketch the form of these densities.

**3.3.5.1 Updating the  $\alpha'$ s** with subscripts 1 for trait 1, 2 for trait 2 and  $k$  for marker set, from the bivariate normal distribution

$$\begin{bmatrix} \alpha_{ki1} \\ \alpha_{ki2} \end{bmatrix} | All, y \sim N \left( \begin{bmatrix} \hat{\alpha}_{ki1} \\ \hat{\alpha}_{ki2} \end{bmatrix}, \left[ I + \frac{\sigma_e^2}{\lambda_{ki}} \begin{pmatrix} \sigma_{g_1}^2 & \sigma_{g_1 g_2} \\ \sigma_{g_1 g_2} & \sigma_{g_2}^2 \end{pmatrix}_k \right]^{-1} \sigma_e^2 \right), \quad i = 1, 2, \dots, n-1,$$

where

$$(I + \lambda_{ki}^{-1} V_{g_k}^{-1} \sigma_e^2) \hat{\alpha}_{is} = \begin{bmatrix} U'_{ki} (y_1 - 1\mu_1 - \sum_{i \neq k} U_i \alpha_{i1}) \\ U'_{ki} (y_2 - 1\mu_2 - \sum_{i \neq k} U_i \alpha_{i2}) \end{bmatrix}, \quad \alpha'_{is} = (\alpha_{i1}, \alpha_{i2}).$$

The strategy updates jointly the  $\alpha'$ s for both traits from a given marker set, conditional on the  $\alpha'$ s from the remaining marker sets.

**3.3.5.2 Updating  $\mu_1$  and  $\mu_2$**  The prior distribution of both scalars  $\mu_1$  and  $\mu_2$  are the normal process  $N(0, 10^5)$ . The fully conditional for  $\mu_1$  is proportional to

$$[\mu_1 | All, x] \sim N \left( \hat{\mu}_1, (1'1 + k_{\mu_1})^{-1} \sigma_e^2 \right)$$

where

$$(1'1 + k_{\mu_1}) \hat{\mu}_1 = 1' (y_m - U \alpha_m - T \gamma_1), \quad k_{\mu_1} = \frac{\sigma_e^2}{10^5}.$$

A similar expression holds for  $\mu_2$ .

**3.3.5.3 Updating  $V_g$**  For each of the marker sets, the update involves drawing samples from scaled inverse Wishart distributions. The fully conditional of the  $2 \times 2$  matrix  $V_g$  is proportional to

$$[\alpha_m, \alpha_f | V_g, D] [V_g | \nu_g, P_g]. \quad (65)$$

The second term is inverse Wishart the prior distribution of  $V_g$ ,  $IW(\nu_g, P_g)$  with density

$$p(V_g | \nu_g, P_g) \propto |V_g|^{-\frac{1}{2}(\nu_g + 3)} \exp \left[ -\frac{1}{2} \text{tr}(V_g^{-1} P_g) \right] \quad (66)$$

where the hyperparameters  $\nu_g$  and  $P_g$  are the degrees of freedom and the scale, respectively. The modal value of this distribution is given by  $P_g/(\nu_g + p + 1)$ , where in our case,  $p = 2$ . On defining (see ?, page 574)

$$S_g = \begin{bmatrix} \alpha'_1 D^{-1} \alpha_1 & \alpha'_1 D^{-1} \alpha_2 \\ \alpha'_2 D^{-1} \alpha_1 & \alpha'_2 D^{-1} \alpha_2 \end{bmatrix},$$

the density of the fully conditional posterior distribution of  $V_g$  is

$$\begin{aligned}
p(V_g|All, y) &\propto |V_g|^{-\frac{k}{2}} |V_g|^{-\frac{1}{2}(v_g+3)} \exp \left[ -\frac{1}{2} \text{tr} (V_g^{-1} P_g) \right] \exp \left[ -\frac{1}{2} \text{tr} (V_g^{-1} S_g) \right] \\
&= |V_g|^{-\frac{1}{2}(k+v_g+3)} \exp \left[ -\frac{1}{2} \text{tr} [V_g^{-1} (S_g + P_g)] \right]
\end{aligned} \tag{67}$$

where  $k = n - 1$ , which is in the form of an inverse Wishart distribution of dimension 2,  $k + v_g$  degrees of freedom and scale matrix  $(S_g + P_g)$ .

**3.3.5.4 Updating  $V_e$**  The update here is similar and it involves again drawing samples from scaled inverse Wishart distributions. The fully conditional posterior distribution of the  $2 \times 2$  matrix  $V_e$  defined in (??) is proportional to

$$[e_1, e_2 | V_e, E] [V_e | \nu_e, P_e] \tag{68}$$

where the second term is the prior distribution of  $V_e$  with hyperparameters  $\nu_e$  and  $P_e$  of the form

$$p(V_e | \nu_e, P_e) \propto |V_e|^{-\frac{1}{2}(v_e+3)} \exp \left[ -\frac{1}{2} \text{tr} (V_e^{-1} P_e) \right] \tag{69}$$

symbolised  $IW(\nu_e, P_e)$ .

The density is

$$p(V_e | All, y) \propto |V_e|^{-\frac{1}{2}(n+v_e+3)} \exp \left[ -\frac{1}{2} \text{tr} [V_e^{-1} (S_e + P_e)] \right],$$

an inverse Wishart distribution of dimension 2,  $n + v_e$  degrees of freedom and scale matrix  $(S_e + P_e)$ , where

$$S_h = \begin{bmatrix} \gamma'_m E^{-1} \gamma_m & \gamma'_m E^{-1} \gamma_f \\ \gamma'_f E^{-1} \gamma_m & \gamma'_f E^{-1} \gamma_f \end{bmatrix}.$$

## References