A multiple-trait Bayesian Lasso for genome-enabled analysis and prediction of complex traits

Daniel Gianola[a,b,c,d] and Rohan L. Fernando[c]

[a] Department of Animal Sciences, University of Wisconsin-Madison, USA;

[b] Department of Dairy Science, University of Wisconsin-Madison, USA;

[c]Department of Animal Science, Iowa State University, USA;

[d] Department of Plant Sciences, Technical University of Munich,

TUM School of Life Sciences, Germany.

[1]Corresponding author. Email: gianola@ansci.wisc.edu

November 20, 2019

1 Abstract

A multiple-trait Bayesian LASSO (MBL) for genome-based analysis and prediction of quantitative traits is presented and applied to two real data sets. The data-generating model is a multivariate linear Bayesian regression on possibly a huge number of molecular markers, and with a Gaussian residual distribution posed. Each (one per marker) of the $T \times 1$ vectors of regression coefficients (T : number of traits) is assigned the same $T-$variate Laplace prior distribution, with a null mean vector and unknown scale matrix $\mathbf{\Sigma}$. The multivariate prior reduces to that of the standard univariate Bayesian LASSO when $T = 1$. The covariance matrix of the residual distribution is assigned a multivariate Jeffreys prior and $\mathbf{\Sigma}$ is given an inverse-Wishart prior. The unknown quantities in the model are learned using a Markov chain Monte Carlo sampling scheme constructed using a scale-mixture of normal distributions representation. MBL is demonstrated in a bivariate context employing two publicly available data sets using a bivariate genomic best linear unbiased prediction model (GBLUP) for benchmarking results. The first data set is one where wheat grain yields in two different environments are treated as distinct traits. The second data set comes from genotyped *Pinus* trees with each individual was measured for two traits, rust bin and gall volume. In MBL, the bivariate marker effects are shrunk

differentially, i.e., "short" vectors are more strongly shrunk towards the origin than in GBLUP; conversely, "long" vectors are shrunk less. A predictive comparison was carried out as well where, in wheat, the comparators of MBL where bivariate GBLUP and bivariate Bayes Cπ, a variable selection procedure. A training-testing layout was used, with 100 random reconstructions of training and testing sets. For the wheat data, all methods produced similar predictions. In *Pinus,* MBL gave better predictions that either a Bayesian bivariate GBLUP or the single trait Bayesian LASSO. MBL has been implemented in the Julia language package JWAS and is now available for the scientific community to explore with different traits, species and environments. It is well known that there is no universally best prediction machine and MBL represents a new piece in the armamentarium for genome-enabled analysis and prediction of complex traits.

2 Introduction

Two main paradigms have been employed for investigating statistical associations between molecular markers and complex traits: marker-by-marker genome-wide association studies (GWAS) and whole-genome regression approaches (WGR). GWAS is dominant in human genetics; Visscher et al. (2017) present a perspective and Gianola et al. (2016) formulate a statistically orientated critique. WGR was developed mostly in animal and plant breeding (e.g., Lande and Thompson 1990; Meuwissen et al. 2001; Gianola et al. 2003) primarily for predicting future performance, but it has received some attention in human genetics as well (e.g., Lee et al. 2011; Yang et al. 2010; de los Campos et al. 2011; Makowsky et al. 2011; López de Maturana et al. 2014). de los Campos et al. (2013), Gianola (2013) and Isik et al. (2017) reviewed an extensive collection of WGR approaches. Other studies noted that WGR can be used both for "discovery" of associations and for prediction (Moser et al. 2015; Goddard et al. 2016; Fernando et al. 2017). Hence, WGR methodology is an active area of research.

Multiple-trait analysis has been of great interest in plant and animal breeding for a long-time, mainly from the point of view of joint selection for many traits (Smith 1936; Hazel 1943; Walsh and Lynch 2018). Henderson and Quaas (1976) developed multi-trait best linear unbiased prediction of breeding values for all individuals and traits measured in a population of animals, a methodology that gradually became routine in the field. For example, Gao et al. (2018), described an application of a 9-variate model to data representing close to seven million and four million Holstein and Nordic Red cattle, respectively; the nine traits were milk, fat and protein yields in each of the first three lactations of the cows.

A multiple-trait analysis is also a natural choice in quests for understanding and dissecting genetic correlations between traits using molecular markers, e.g., evaluating whether pleiotropy or linkage disequilibrium are at the roots of between-trait associations (Gianola et al. 2015; Cheng et al. 2018a). For instance, Galesloot et al. (2014) compared six methods of multivariate

GWAS via simulation and found that all delivered a higher power than single-trait GWAS, even when genetic correlations were weak. Many single-trait WGR methods extend directly to the multiple-trait domain, e.g., genomic best linear unbiased prediction (GBLUP; Van Raden 2007). Other procedures such as Bayesian mixture models are more involved, but extensions are available (Calus and Veerkamp 2011; Jia and Jannink 2012; Cheng et al. 2018a). The mixture model of Cheng et al. (2018a) 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 $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ categories, where $(0,0)$ means "no effect" and $(1,1)$ denotes "effect" on each of two disease traits in *Pinus taeda* was estimated by Cheng et al. (2018a) using SNPs (single nucleotide polymorphisms). The proportion of MCMC samples falling into the $(1,1)$ class was less than 3%, with about 140 markers appearing as candidates for further scrutiny of pleiotropy; 97% of the SNP were in the $(0,0)$ class and 0.5% were in the $(0,1)$ and $(1,0)$ classes. It must be noted that Cheng et al. (2018a) used Bayesian model averaging, so posterior estimates of effects and of their uncertainties constitute averages over all possible configurations. The resulting "average model" is not truly sparse as Bayesian mixture models always assign some posterior probability to each of the possible configurations. An alternative to a mixture is to use a prior distribution that produces strong shrinkage towards the origin of "weak" vectors of marker effects; here, each marker has a vector with dimension equal to the number of traits.

The LASSO (least absolute shrinkage and selection operator) presented by Tibshirani (1996) is a single-response method based on minimizing a linear regression residual sum of squares subject to a constraint based on an L_1 norm. It can produce a sparse model, i.e., if the linear regression model has p regression coefficients, the LASSO yields a smaller model (i.e., model selection) but with a complexity that cannot exceed N, the number of observations. Tibshirani (1996) noted that the LASSO solutions can also be obtained by calculating the mode of the conditional posterior distribution of the regression coefficients in a Bayesian model in which each coefficient is assigned the same conditional double exponential or Laplace prior. Using a ridge regression reformulation of LASSO, it can be seen (Tibshirani 1996; Gianola 2013) that its Bayesian version shrinks small-value regression coefficients very strongly towards zero, whereas large-effect variants are regularized to a much lesser extent than in ridge regression. Yuan and Lin (2006) and Yuan et al. (2007) considered the problem of clustering regression coefficients into groups (factors), with the focus becoming factor selection, as opposed to the predictor variable selection that takes place in LASSO. For instance, a cluster could consist of a group of markers in tight physical linkage. These authors noted that, in some instances, grouping enhances prediction performance over ridge regression, while in others, it does not. Such finding is consisting with knowledge accumulated in close to two decades of experience with genome-enabled prediction in animal breeding: there is no universally best prediction machine. A multiple-trait application

of a LASSO penalty on regression coefficients was presented by Li et al. (2015). These authors assigned a multivariate Laplace distribution to the model residuals and a group-LASSO penalty (Yuan and Lin 2006) to the regression coefficients. The procedure differs from Tibshirani's LASSO in that the model selects vectors of regressors (corresponding to regressions of a given marker over traits) as opposed to single-trait predictor variables.

Park and Casella (2008) introduced a fully Bayesian LASSO (BL). Contrary to LASSO, BL produces a model where all regression coefficients are non-null (even if $p > N$); most regressions are often tiny in value, except those associated with covariates (markers) with strong effects. In short, LASSO produces a sparse model whereas BL yields an effectively sparse specification, similar to Bayesian mixture models such as Bayes B (Meuwissen et al. 2001). The fist application of the BL in quantitative genetics was made by Yi and Xu (2008) in the context of quantitative trait locus (QTL) mapping, with subsequent applications in de los Campos et al. (2009), Legarra et al. (2011), Li et al. (2011) and Lehermeier et al. (2013).

It appears that a multiple-trait generalization of the BL has not been reported hereto. The present paper describes a multi-trait Bayesian LASSO (MBL) model based on adopting a multivariate Laplace distribution with unknown scale matrix as prior distribution for the markers or variants under scrutiny. The MBL is introduced and compared with a multiple-trait GBLUP (MTGBLUP) using wheat and pine tree data sets. Section "The multi-trait regression model" describes MBL, including a Markov chain Monte Carlo sampling algorithm. Subsequently, MBL is compared with MTGBLUP using a wheat data set. Finally, bivariate MBL and bivariate MTGBLUP are contrasted from a predictive perspective, showing a better performance of MBL over BLUP and over a single-trait Bayesian LASSO specification, corroborating the usefulness of multiple-trait analyses. The paper concludes with a general discussion and technical details are presented in Appendices.

3 The multi-trait regression model

Assume there are T traits observed in each of N individuals and let $\boldsymbol{\beta}_j = \{\beta_{jt}\}$ be a $T \times 1$ vector of allelic substitution effects at marker $j = 1, 2, ..., p$, with β_{jt} representing the effect of marker j on trait t ($t = 1, 2, ..., T$). The multi-trait regression model (assuming no nuisance location effects other than a mean) for the T responses is

$$\mathbf{y}_i = \boldsymbol{\mu} + \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j + \mathbf{e}_i; \;\; i = 1, 2..., N; \;\; j = 1, 2, ..., p, \tag{1}$$

where $\mathbf{y}_i$ is a $T \times 1$ vector of responses observed in individual i; $\boldsymbol{\mu} = \{\mu_t\}$ is the vector of trait means and x_{ij} is the genotype individual i possesses at marker locus j. The residual vector

$\mathbf{e}_i$ $(T \times 1)$ is assumed to follow the Gaussian distribution $\mathbf{e}_i|\mathbf{R}_0 \sim N(0, \mathbf{R}_0)$, where $\mathbf{R}_0$ is a $T \times T$ covariance matrix. All $\mathbf{e}_i$ vectors are assumed to be mutually independent and identically distributed.

If traits are sorted within individuals, the probability model associated with (1) can be represented as

$$\begin{aligned} & p\left(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N | \boldsymbol{\mu}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p, \mathbf{R}_0\right) \\ \propto & \frac{1}{|\mathbf{R}_0|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j\right)' \mathbf{R}_0^{-1}\left(\mathbf{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j\right)\right] \\ \propto & \frac{1}{|\mathbf{R}_0|^{\frac{N}{2}}} \exp\left\{-\frac{1}{2} tr\left[\mathbf{R}_0^{-1}\mathbf{S}_e\right]\right\}, \end{aligned} \tag{2}$$

where

$$\mathbf{S}_e = \sum_{i=1}^{N}\left(\mathbf{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j\right)\left(\mathbf{y}_i - \boldsymbol{\mu} - \sum_{j=1}^{p} x_{ij}\boldsymbol{\beta}_j\right)' \tag{3}$$

is a $T \times T$ matrix of sums of squares and products of the unobserved regression residuals.

The regression model can be formulated in an equivalent manner by sorting individuals within traits; we will use $T = 3$ hereinafter. Let $\mathbf{y}_1^*$, $\mathbf{y}_2^*$ and $\mathbf{y}_3^*$ be response vectors of order N each observed for traits 1, 2, and 3, respectively. The representation of the model is

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_1^* \\ \mathbf{y}_2^* \\ \mathbf{y}_3^* \end{bmatrix} &= \begin{bmatrix} \mathbf{1}_N & 0 & 0 \\ 0 & \mathbf{1}_N & 0 \\ 0 & 0 & \mathbf{1}_N \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \mathbf{X} & 0 & 0 \\ 0 & \mathbf{X} & 0 \\ 0 & 0 & \mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \\ \boldsymbol{\beta}_3^* \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1^* \\ \mathbf{e}_2^* \\ \mathbf{e}_3^* \end{bmatrix} \\ &= (\mathbf{I}_3 \otimes \mathbf{1}_N)\boldsymbol{\mu} + (\mathbf{I}_3 \otimes \mathbf{X})\boldsymbol{\beta}^* + \mathbf{e}^*, \end{aligned} \tag{4}$$

where $\mathbf{1}_N$ is an $N \times 1$ vector of $1's$, $\mathbf{X} = \{x_{ij}\}$ is an $N \times p$ matrix of marker genotypes, and $\boldsymbol{\beta}_t^*$ $(p \times 1)$ and $\mathbf{e}_t^*$ $(N \times 1)$ are vectors of regression coefficients and of residuals for trait t, respectively. Above, $\boldsymbol{\beta}^* = vec\left(\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*, \boldsymbol{\beta}_3^*\right)$ is a $3p \times 1$ vector and $\mathbf{e}^* = vec(\mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{e}_3^*)$ has dimension $3N \times 1$. Note that $Var\left(\mathbf{e}^*\right) = \mathbf{R}_0 \otimes \mathbf{I} = \mathbf{R}$. Putting $\overline{\mathbf{y}}^*\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*\right) = (\mathbf{I}_3 \otimes \mathbf{1}_N)\boldsymbol{\mu} + (\mathbf{I}_3 \otimes \mathbf{X})\boldsymbol{\beta}^*$, the probability model is

$$p\left(\mathbf{y}^* | \boldsymbol{\mu}, \boldsymbol{\beta}^*, \mathbf{R}_0\right) \propto \frac{1}{|\mathbf{R}_0|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{y}^* - \overline{\mathbf{y}}^*\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*\right)\right)' \mathbf{R}^{-1}\left(\mathbf{y}^* - \overline{\mathbf{y}}^*\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*\right)\right)\right]. \tag{5}$$

We will work with either (1) or (4), depending on the context.

3.1 Bayesian prior assumptions

3.1.1 Parameters $\boldsymbol{\mu}$ and $\mathbf{R}_0$

The vector $\boldsymbol{\mu}$ will be assigned a "flat" improper prior and Jeffreys non-informative prior (e.g., Sorensen and Gianola, 2002) will be adopted for $\mathbf{R}_0$ so that their joint prior density is

$$p(\boldsymbol{\mu}, \mathbf{R}_0) \propto |\mathbf{R}_0|^{-\left(\frac{T+1}{2}\right)}. \tag{6}$$

3.1.2 Multivariate Laplace prior distribution (MLAP) for marker effects

The same T-variate Laplace prior distribution with a null mean vector will be assigned to each of the $T \times 1$ vectors $\boldsymbol{\beta}_j$ $(j = 1, 2, ..., p)$, assumed mutually independent, *a priori*. Gómez et al. (2007) presented a multi-dimensional version of the power exponential family of distributions; one special case is the multivariate Laplace distribution (MLAP). The density of the MLAP with a zero-mean vector used here is

$$p(\boldsymbol{\beta}_j|\boldsymbol{\Sigma}) = \frac{T\Gamma\left(\frac{T}{2}\right)}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}\,\pi^{\frac{T}{2}}\Gamma\left(1+T\right)2^{(1+T)}} \exp\left(-\frac{1}{2}\sqrt{\boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j}\right);\ j = 1, 2, ..., p, \tag{7}$$

where $\boldsymbol{\Sigma} = \{\Sigma_{tt'}\}$ is a $T \times T$ positive-definite scale matrix. The variance-covariance matrix of MLAP is

$$Var(\boldsymbol{\beta}_j|\boldsymbol{\Sigma}) = 4\left(T+1\right)\boldsymbol{\Sigma} = \mathbf{B}; \tag{8}$$

note that the absolute values of the elements of $\mathbf{B}$, the inter-trait variance-covariance of marker effects, are larger than those of $\boldsymbol{\Sigma}$. Hence, $\beta_{jt} \sim \left(0, \sigma^2_{\beta,t}\right)$ for $\forall j$, where $\sigma^2_{\beta,t} = 4\left(T+1\right)\Sigma_{tt}$ is the appropriate diagonal element of $\mathbf{B}$; likewise, $\sigma_{\beta,tt'} = 4\left(T+1\right)\Sigma_{tt'}$ is the covariance of marker effects between traits t and t', for all j. Putting $T = 1$ in (7) yields

$$p(\beta|\Sigma) = \frac{1}{2\sqrt{4\Sigma}} \exp\left(-\frac{|\beta|}{\sqrt{4\Sigma}}\right). \tag{9}$$

The preceding is the density of a double exponential (DE) distribution with null mean, parameter $\sqrt{4\Sigma}$ and variance $Var(\beta) = 8\Sigma$. As mentioned earlier, Tibshirani (1996) and Park and Casella (2008) used the DE distribution as conditional (given Σ) prior for regression coefficients in the BL, a member of the "Bayesian Alphabet" (Gianola et al. 2009). Gianola et al. (2018) assigned the DE distribution to residuals of a linear model for the purpose of attenuating outliers and Li et al. (2015) used the MLAP distribution for the residuals in a "robust" linear regression model for QTL mapping.

MLAP is therefore an interesting candidate prior for multi-trait marker effects in a multiple trait generalization of the Bayesian LASSO (MBL). A zero-mean MLAP distribution has a

sharp peak at the 0 coordinates. Although when $T = 1$ MLAP reduces to a DE distribution, the marginal and conditional densities of MLAP are not DE. Gómez et al. (2007) showed that such densities are elliptically contoured, and thus not DE. Appendix A and Figures S1-S3 in the Supplemental material give background on MLAP.

Gómez-Sánchez-Manzano et al. (2008) showed that MLAP can be represented as a scaled mixture of normal distributions under the hierarchy: 1) $\left[\boldsymbol{\beta}_j|\boldsymbol{\Sigma},v_j^2\right] = N_T\left(\mathbf{0},v_j^2\boldsymbol{\Sigma}\right)$, and 2) $v_j^2 \sim Gamma\left(\frac{T+1}{2},\frac{1}{8}\right)$ for $j = 1,2,...,p$; $N_T\left(\mathbf{0},\boldsymbol{\Sigma}v_j^2\right)$ denotes a $T-$variate normal distribution with null mean and covariance matrix $\boldsymbol{\Sigma}v_j^2$. The density of v_j^2 is

$$h\left(v_j^2\right) \propto \left(v_j^2\right)^{\frac{T+1}{2}-1}\exp(-\frac{v_j^2}{8}) \,. \tag{10}$$

Let the collection of all marker effects over traits be represented by the $Tp \times 1$ vector

$$\boldsymbol{\beta} = \left[\begin{array}{cccc} \boldsymbol{\beta}_1' & \boldsymbol{\beta}_2' & . \; . \; . & \boldsymbol{\beta}_p' \end{array}\right]'. \tag{11}$$

If independent and identical MLAP prior distributions are assigned to each of the sub-vectors, the joint prior density of all marker effects, given $\boldsymbol{\Sigma}$, can be represented as

$$\begin{aligned} p\left(\boldsymbol{\beta}|\boldsymbol{\Sigma}\right) &= \prod_{j=1}^{p} p\left(\boldsymbol{\beta}_j|\boldsymbol{\Sigma}\right) \\ &= \prod_{j=1}^{p}\int_0^\infty N_T\left(\boldsymbol{\beta}_j|0,\boldsymbol{\Sigma}v_j^2\right)h\left(v_j^2\right)dv_j^2, \end{aligned} \tag{12}$$

and the joint density of $\boldsymbol{\beta}$ and $\mathbf{v}^2 = \left[\begin{array}{cccc} v_1^2 & v_2^2 & . \; . \; . & v_p^2 \end{array}\right]'$ is

$$p\left(\boldsymbol{\beta},\mathbf{v}^2|\boldsymbol{\Sigma}\right) = \prod_{j=1}^{p} N_T\left(0,\boldsymbol{\Sigma}v_j^2\right)h\left(v_j^2\right). \tag{13}$$

When individuals are sorted within traits (e.g., $T = 3$), note that $\left[\boldsymbol{\beta}^*|\boldsymbol{\Sigma},\mathbf{v}^2\right]$ is a $Tp-$dimensional normal distribution with null mean vector and covariance matrix

$$Var\left(\left[\begin{array}{c}\boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\\ \boldsymbol{\beta}_3^*\end{array}\right]|\boldsymbol{\Sigma},\mathbf{v}^2\right) = \left[\begin{array}{ccc}\mathbf{D}\Sigma_{11} & \mathbf{D}\Sigma_{12} & \mathbf{D}\Sigma_{13}\\ \mathbf{D}\Sigma_{21} & \mathbf{D}\Sigma_{22} & \mathbf{D}\Sigma_{23}\\ \mathbf{D}\Sigma_{31} & \mathbf{D}\Sigma_{32} & \mathbf{D}\Sigma_{32}\end{array}\right] = \boldsymbol{\Sigma}\otimes\mathbf{D}, \tag{14}$$

where $\mathbf{D} = diag\left(v_1^2, v_2^2, ..., v_p^2\right)$ is a diagonal matri. Hence,

$$p\left(\boldsymbol{\beta}^*|\boldsymbol{\Sigma}, \mathbf{v}^2\right) \propto \exp\left[-\frac{1}{2}\boldsymbol{\beta}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{D}^{-1}\right)\boldsymbol{\beta}^*\right]. \tag{15}$$

3.1.3 Scale matrix $\boldsymbol{\Sigma}$

The scale matrix $\boldsymbol{\Sigma}$ of MLAP can be given a fixed value (becoming a hyper-parameter) or inferred, in which case a prior distribution is needed. Here, an inverse-Wishart (IW) distribution with scale matrix $\boldsymbol{\Omega}_\beta$ and ν_β degrees of freedom will be assigned as prior. The density is

$$p\left(\boldsymbol{\Sigma}|\Omega_\beta, \nu_\beta\right) \propto |\boldsymbol{\Sigma}|^{-\left(\frac{T+\nu_\beta+1}{2}\right)} \exp\left[-\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\Omega_\beta\right)\right]. \tag{16}$$

3.2 Joint posterior and fully-conditional distributions

The joint posterior distribution, including $\mathbf{v}^2 = \left\{v_j^2\right\}$ from the scale-mixture of normals representation of the prior distribution of $\boldsymbol{\beta}$, was assumed to take the form

$$p\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*, \mathbf{R}_0, \boldsymbol{\Sigma}, \mathbf{v}^2|\mathbf{y}^*, H\right) \propto p\left(\mathbf{y}^*|\boldsymbol{\mu}, \boldsymbol{\beta}^*, \mathbf{R}_0\right) p\left(\mathbf{R}_0|H\right) p\left(\boldsymbol{\beta}^*|\boldsymbol{\Sigma}, \mathbf{v}^2\right) p\left(\mathbf{v}^2\right) p\left(\boldsymbol{\Sigma}|H\right), \tag{17}$$

where H denotes the hyper-parameters; recall that $\mathbf{y}^*$ is the data vector sorted by individuals within trait The fully conditional distributions are presented below, with $ELSE$ used to denote all parameters that are kept fixed, together with H, in a specific conditional distribution.

3.2.1 Parameters μ and $\boldsymbol{\beta}^*$ given $ELSE$

From (17) and using representations (4) and (15), the fully conditional posterior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}^*$ has density

$$\begin{aligned} p\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*|ELSE\right) &\propto p\left(\mathbf{y}^*|\boldsymbol{\mu}, \boldsymbol{\beta}^*, \mathbf{R}_0\right) p\left(\boldsymbol{\beta}^*|\boldsymbol{\Sigma}, \mathbf{v}^2\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\mathbf{y}^* - \overline{\mathbf{y}}^*\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*\right)\right)'\mathbf{R}^{-1}\left(\mathbf{y}^* - \overline{\mathbf{y}}^*\left(\boldsymbol{\mu}, \boldsymbol{\beta}^*\right)\right)\right] \\ &\quad \times \exp\left[-\frac{1}{2}\boldsymbol{\beta}^{*\prime}\left(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{D}^{-1}\right)\boldsymbol{\beta}^*\right]. \end{aligned} \tag{18}$$

The preceding is a multivariate normal density (e.g., Sorensen and Gianola 2002). The mean vector of the distribution is

$$\begin{bmatrix} \overline{\boldsymbol{\mu}} \\ \overline{\boldsymbol{\beta}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{R}_0^{-1}N & \mathbf{R}_0^{-1}\otimes\mathbf{1}_N'\mathbf{X} \\ \mathbf{R}_0^{-1}\otimes\mathbf{X}'\mathbf{1}_N & \mathbf{R}_0^{-1}\otimes\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}^{-1}\otimes\mathbf{D}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \left(\mathbf{R}_0^{-1}\otimes\mathbf{1}_N'\right)\mathbf{y}^* \\ \left(\mathbf{R}_0^{-1}\otimes\mathbf{X}'\right)'\mathbf{y}^* \end{bmatrix}, \tag{19}$$

and the variance-covariance matrix is

$$Var\left(\left[\begin{array}{c}\boldsymbol{\mu}\\ \boldsymbol{\beta}\end{array}\right]|ELSE\right)=\left[\begin{array}{cc}\mathbf{R}_0^{-1}N & \mathbf{R}_0^{-1}\otimes\mathbf{1}_N'\mathbf{X}\\ \mathbf{R}_0^{-1}\otimes\mathbf{X}'\mathbf{1}_N & \mathbf{R}_0^{-1}\otimes\mathbf{X}'\mathbf{X}+\boldsymbol{\Sigma}^{-1}\otimes\mathbf{D}^{-1}\end{array}\right]^{-1}=\mathbf{C}^{-1}. \tag{20}$$

A more explicit representation is presented in Appendix B for the case $T=3$.

3.2.2 Fully conditional distributions of partitions of the location vector

For details, see Van Tassell and Van Vleck (1996) and Sorensen and Gianola (2002). Since the joint posterior of the location parameters, given $\boldsymbol{\Sigma}$, $\mathbf{v}^2$ and $\mathbf{R}_0$, is multivariate normal, all conditionals and linear combinations thereof are normal as well. In particular ($T=3$),

$$E\left(\mu_t|ELSE\right)=\frac{1}{r^{tt}N}\left[\mathbf{1}_N'\sum_{t'=1}^{3}r^{tt'}\left(\mathbf{y}_t^*-\mathbf{X}\boldsymbol{\beta}t\right)-N\sum_{t'\neq t}r^{tt'}\mu_{t'}\right];\ i=1,2,3, \tag{21}$$

and

$$Var\left(\mu_t|ELSE\right)=\frac{1}{r^{tt}N};\ t=1,2,3. \tag{22}$$

Likewise

$$E\left(\boldsymbol{\beta}_t^*|ELSE\right)=\left(r^{tt}\mathbf{X}'\mathbf{X}+\frac{\mathbf{D}^{-1}}{\Sigma^{tt}}\right)^{-1}$$
$$\times\left[\mathbf{X}'\sum_{t'=1}^{3}r^{tt'}\left(\mathbf{y}_j^*-\mathbf{1}_N\mu_j\right)-\sum_{t'\neq t}\left(r^{tt'}\mathbf{X}'\mathbf{X}+\frac{\mathbf{D}^{-1}}{\Sigma^{tt'}}\right)\boldsymbol{\beta}_{t'}^*\right];\ i=1,2,3, \tag{23}$$

and

$$Var\left(\boldsymbol{\beta}_t^*|ELSE\right)=\left(r^{tt}\mathbf{X}'\mathbf{X}+\frac{\mathbf{D}^{-1}}{\Sigma^{tt}}\right)^{-1};\ t=1,2,3. \tag{24}$$

3.2.3 Fully conditional distributions of $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$

From (17) using (2) and (6)

$$p\left(\mathbf{R}_0|ELSE\right)\propto|\mathbf{R}_0|^{-\left(\frac{N+T+1}{2}\right)}\exp\left\{-\frac{1}{2}tr\left[\mathbf{R}_0^{-1}\mathbf{S}_e\right]\right\}, \tag{25}$$

so $[\mathbf{R}_0|ELSE]$ is an IW distribution with $N+T$ degrees of freedom and scale matrix $\mathbf{S}_e$. In IW, the kernel of the density is often written as $\exp\left\{-\frac{1}{2}tr\left[\mathbf{R}_0^{-1}\left(N+T\right)\overline{\mathbf{S}}_e\right]\right\}$, where $\overline{\mathbf{S}}_e=\mathbf{S}_e/\left(N+T\right)$.

Recall that

$$p\left(\boldsymbol{\beta}^{*}|\boldsymbol{\Sigma},\mathbf{v}^{2}\right)=\prod_{j=1}^{p}N_{T}\left(\boldsymbol{\beta}_{j}|0,\boldsymbol{\Sigma}v_{j}^{2}\right), \tag{26}$$

so from (17)

$$\begin{aligned} p\left(\boldsymbol{\Sigma}|ELSE\right) &\propto p\left(\boldsymbol{\beta}^{*}|\boldsymbol{\Sigma},\mathbf{v}^{2}\right)p\left(\boldsymbol{\Sigma}|H\right) \\ &\propto \prod_{j=1}^{p}\frac{1}{\left|\boldsymbol{\Sigma}v_{j}^{2}\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\boldsymbol{\beta}_{j}'\left(\frac{\boldsymbol{\Sigma}^{-1}}{v_{j}^{2}}\right)\boldsymbol{\beta}_{j}\right] \\ &\quad\times\left|\boldsymbol{\Sigma}\right|^{-\left(\frac{T+\nu_{\beta}+1}{2}\right)}\exp\left[-\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{-1}\Omega_{\beta}\right)\right] \\ &\propto \left|\boldsymbol{\Sigma}\right|^{-\left(\frac{p+T+\nu_{\beta}+1}{2}\right)}\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}_{\beta}\right]\right\}, \end{aligned} \tag{27}$$

where

$$\mathbf{S}_{\beta}=\sum_{j=1}^{p}\left(\frac{\boldsymbol{\beta}_{j}\boldsymbol{\beta}_{j}'}{v_{j}^{2}}\right)+\Omega_{\beta} \tag{28}$$

is a $T\times T$ matrix. Hence the conditional posterior distribution of $\boldsymbol{\Sigma}$ is $IW\left(p+T+\nu_{\beta},\mathbf{S}_{\beta}\right)$. The kernel of the density of $\boldsymbol{\Sigma}$ is often represented as $\exp\left\{-\frac{1}{2}tr\left[\boldsymbol{\Sigma}_{0}^{-1}\left(p+T+\nu_{\beta}\right)\overline{\mathbf{S}}_{\beta}\right]\right\}$, where $\overline{\mathbf{S}}_{\beta}=\mathbf{S}_{\beta}/\left(p+T+\nu_{\beta}\right)$.

3.2.4 Fully conditional distribution of $\mathbf{v}^2$

From (17) and using (13)

$$\begin{aligned} p\left(\mathbf{v}^{2}|ELSE\right) &\propto p\left(\boldsymbol{\beta}^{*}|\boldsymbol{\Sigma},\mathbf{v}^{2}\right)p\left(\mathbf{v}^{2}\right) \\ &\propto \prod_{j=1}^{p}N_{T}\left(\boldsymbol{\beta}_{j}|0,\boldsymbol{\Sigma}v_{j}^{2}\right)h(v_{j}^{2}) \\ &\propto \prod_{j=1}^{p}\frac{1}{\left(v_{j}^{2}\right)^{\frac{T}{2}}}\exp\left[-\frac{\boldsymbol{\beta}_{j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_{j}}{2v_{j}^{2}}\right]\left(v_{j}^{2}\right)^{\frac{T+1}{2}-1}\exp(-\frac{v_{j}^{2}}{8}) \\ &\propto \prod_{j=1}^{p}\left(v_{j}^{2}\right)^{-\frac{1}{2}}\exp\left[-\frac{\boldsymbol{\beta}_{j}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_{j}+\frac{v^{4}}{4}}{2v_{j}^{2}}\right]. \end{aligned} \tag{29}$$

The preceding density is not in a recognizable form. Appendix C gives details of a Metropolis-Hastings algorithm tailored for making draws from the distribution having density (29). A brief description of the procedure follows.

3.3 MCMC algorithm

- Starting values for $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$ can be obtained "externally" from some estimates of $\mathbf{R}_0$ and $\mathbf{B}$ (the $T \times T$ matrix of variances and covariances of marker effects) calculated with standard methods such as maximum likelihood. Recall that $\boldsymbol{\Sigma} = \mathbf{B}/[4(T+1)]$.

- Sample each v_j^2 $(j = 1, 2, ..., p)$ using the following Metropolis-Hastings sampler:

  1. At round t, draw y from $Y \sim IG(\alpha = \frac{1}{2}, \beta = \frac{1}{4})$ and evaluate y as proposal; IG stands for an inverse-gamma distribution.
  2. Draw $U \sim U(0,1)$, with the probability of move being $\min(1, R)$, with R as in Appendix C.
  3. If $U < R$, set $w_j^{[t+1]} = y$ and form $v_j^{2[t+1]} = 2/w_j^{[t]}$ as a new state. Otherwise, set $v_j^{2[t+1]} = v_j^{2[t]}$; $j = 1, 2, ..., p$.

- In a "single-pass" sampler, use (19) and (20) for sampling the entire location vector jointly. Otherwise, adopt a blocking strategy; for example draw $\boldsymbol{\mu}$ and $\boldsymbol{\beta}^*$ using (21), (22), (23) and (24).

- Sample $\mathbf{R}_0$ from $IW(N + T, \mathbf{S}_e)$ and $\boldsymbol{\Sigma}$ from $IW(p + T + \nu_\beta, \mathbf{S}_\beta)$.

3.4 Remarks

Appendix E shows that the degree of shrinkage of marker effects results from a joint action between $\boldsymbol{\Sigma}$ and the strength of marker effects. A vector of effects of a marker with a short Mahalanobis distance away from $\mathbf{0}$ is more strongly shrunk towards the origin (i.e., the mean of prior distribution) than vectors containing strong effects on at least one trait. MLAP preserves the spirit of BL, producing "pseudo-selection" of covariates: all markers stay in the model, but some are effectively nullified. A marker with strong marginal and joint effects on the traits under consideration could flag potentially pleiotropic regions.

3.5 Missing records for some traits

Often, not all traits are measured in all individuals, a situation that is more common in animal breeding than in plant breeding. A standard approach ("data augmentation") treats missing phenotypes as unknowns in an expanded joint posterior distribution. As shown in Appendix F, a predictive distribution can be used to produce an imputation of missing data.

4 Alternative formulation in TN dimensions

The MCMC sampler described above is based on a regression on markers formulation stemming from either (1) or (4). In a "single-pass" sampler, $T(1+p)$ parameters must be drawn together; when p is very large, direct inversion is typically unfeasible so the scheme must be reformulated into a "block-sampling" one, i.e., by drawing some of the location parameters jointly by conditioning on the other location parameters, or by using a single-site sampler (Sorensen and Gianola 2002). Blocking or single-site sampling facilitate computation at the expense of slowing down convergence to the target distribution. Appendix D gives a scheme in which $T(1+N)$ effects (trait means and bivariate genomic breeding values) are inferred, and the Tp marker effects are calculated indirectly, following ideas of Henderson (1977) and adapted by Goddard (2009) to a genome-based model.

5 Data availability statement

The wheat yield data set in the `R` package BGLR (Pérez and de los Campos 2014) was employed to contrast MBL with GBLUP and Bayes Cπ. This wheat data set has been studied extensively, e.g., by Crossa et al. (2010), Gianola et al. (2011), Long et al. (2011) and Gianola et al. (2016). There are $n = 599$ wheat inbred lines, each genotyped with $p = 1279$ DArT (Diversity Array Technology) markers and each planted in four environments. The DArT markers are binary $(0, 1)$ and denote presence or absence of an allele at a marker locus in a given line. Grain yields in environments 1 and 2 were employed to compare outcomes between analyses based on bivariate GBLUP and the bivariate BL. In the bivariate model, yields in the two environments are treated as distinct traits, conceptually, an idea that dates back to Falconer (1952). This type of setting arises in dairy cattle-breeding, where milk production of daughters of bulls in different countries are regarded as different traits and in multi-environment situations in plant breeding; both instances can be represented as special cases of a multiple-trait mixed effects model.

A publicly available Loblolly pine (*Pinus taeda*) data described in Cheng et al. (2018a) was used to carry out a predictive comparison between a Bayesian bivariate GBLUP with the bivariate Bayesian LASSO, as well as the latter versus a single-trait Bayesian LASSO. After edits, there were $n = 807$ individuals with $p = 4828$ SNP markers with measurements on rust bin scores and rust gall volume, two disease traits; see Cheng et al. (2018a).

6 Bivariate analysis of wheat yield: MBL versus GBLUP

6.1 Genomic BLUP and Bayesian BLUP

The bivariate model was

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \end{bmatrix} + \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \tag{30}$$

where $\mathbf{y}_1$ ($\mathbf{y}_2$) is the vector of grain yields in environment 1 (2) of the 599 inbred lines; μ_1 and μ_2 are the trait means in the two environments and $\mathbf{1}$ is a 599×1 incidence vector of ones; $\mathbf{g}_1$ and $\mathbf{g}_2$ are the "additive genomic values" of the lines and $\mathbf{e}_1$ and $\mathbf{e}_2$ are model residuals. In GBLUP (Van Raden 2008) the genetic signals captured by markers are represented as $\mathbf{g}_1 = \mathbf{X}\boldsymbol{\beta}_1$ and $\mathbf{g}_2 = \mathbf{X}\boldsymbol{\beta}_2$ where $\mathbf{X}$ is a 599×1279 centered and scaled matrix of genotype codes, and $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_2$) contains the marker allele substitution effects on trait 1 (2). The residual distribution was

$$\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{R}_0 \otimes \mathbf{I}\right), \tag{31}$$

where, as-before, $\mathbf{R}_0$ is the 2×2 between-trait residual variance-covariance matrix. Effects of environment 1 are expected to be uncorrelated with those of environment 2. However, allowance was made for a non-null residual covariance because the additive genomic model may not capture extant epistasis involving additive effects, potentially creating correlations among residuals of the same lines in different environmental conditions.

GBLUP assumed $\boldsymbol{\beta}_1 \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_{\beta_1}\right)$, $\boldsymbol{\beta}_2 \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_{\beta_2}\right)$ and $Cov\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2'\right) = \mathbf{I}\sigma_{\beta_1\beta_2}$, so

$$\mathbf{B} = \begin{bmatrix} \sigma^2_{\beta_1} & \sigma_{\beta_1\beta_2} \\ \sigma_{\beta_1\beta_2} & \sigma^2_{\beta_2} \end{bmatrix} \tag{32}$$

is the variance-covariance matrix of marker effects. It follows that

$$\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{G}_0 \otimes \mathbf{G}\right), \tag{33}$$

where

$$\mathbf{G}_0 = p\mathbf{B} = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} \\ \sigma_{g_{12}} & \sigma^2_{g_2} \end{bmatrix}, \tag{34}$$

is a between-trait variance-covariance matrix of the additive genomic values (here, e.g., $\sigma^2_{g_1} = p\sigma^2_{\beta_1}$) and $\mathbf{G} = \mathbf{X}\mathbf{X}'/p$ is a genomic-relationship matrix describing genome-based similarities

among the 599 lines. The preceding assumptions induce the marginal distribution

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \sim \mathbf{N}\left(\begin{bmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \end{bmatrix}, \mathbf{V} = \mathbf{G}_0 \otimes \mathbf{G} + \mathbf{R}_0 \otimes \mathbf{I} \right), \tag{35}$$

where $\mathbf{V}$ is the phenotypic covariance-matrix. The bivariate best linear unbiased predictor of $\mathbf{g}_1$ and $\mathbf{g}_2$ (Henderson 1975)

$$\begin{bmatrix} \widehat{\mathbf{g}}_1 \\ \widehat{\mathbf{g}}_2 \end{bmatrix} = (\mathbf{G}_0 \otimes \mathbf{G})\, \mathbf{V}^{-1} \left(\begin{bmatrix} \mathbf{y}_1 - \mathbf{1}\widehat{\mu}_1 \\ \mathbf{y}_2 - \mathbf{1}\widehat{\mu}_2 \end{bmatrix} \right), \tag{36}$$

where

$$\begin{bmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \end{bmatrix} = \left(\begin{bmatrix} \mathbf{1}' & \mathbf{0} \\ \mathbf{0} & \mathbf{1}' \end{bmatrix} \mathbf{V}^{-1} \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} \mathbf{1}' & \mathbf{0} \\ \mathbf{0} & \mathbf{1}' \end{bmatrix} \mathbf{V}^{-1} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right), \tag{37}$$

is the bivariate generalized least-squares (GLS) estimator of the trait means.

BLUP and GLS require knowledge of $\mathbf{G}_0$ and $\mathbf{R}_0$ and we replaced these unknown matrices by estimates obtained using a crude but simple procedure. Genomic and residual variance components were obtained by univariate maximum likelihood analyses of traits 1, 2 and $1+2$, and covariance component estimates were calculated from the expression $Cov(X,Y) = [Var(X+Y) - Var(X) - Var(Y)]/2$. The resulting estimates of $\mathbf{G}_0$ and $\mathbf{R}_0$ were inside of their corresponding parameter spaces. An estimate of $\mathbf{B}$ was obtained by applying relationship (34) to the estimated $\mathbf{G}_0$.

Henderson (1977) showed how BLUP of vectors that are not likelihood identified can be obtained from best linear unbiased predictions of likelihood-identified random effects (see Gianola 2013). Goddard (2009) and Strandén and Garrick (2009) used this property to obtain predictions of marker effects ($\boldsymbol{\beta}$) given predictions of signal ($\mathbf{g}$). If $\boldsymbol{\beta}$ and $\mathbf{g}$ have a joint normal distribution, under (30) one has

$$E\left(\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} | \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} \right) = \left(\mathbf{B}\mathbf{G}_0^{-1} \otimes \mathbf{X}'\mathbf{G}^{-1}\right) \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix}.$$

Using iterated expectations and recalling that BLUP can be viewed as an estimated conditional expectation (with fixed effects replaced by their GLS estimates), BLUP of marker effects is expressible as

$$\begin{aligned} \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} &= \widehat{E}\left(\begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} | \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right) = \left(\mathbf{B}\mathbf{G}_0^{-1} \otimes \mathbf{X}'\mathbf{G}^{-1}\right) \begin{bmatrix} \widehat{\mathbf{g}}_1 \\ \widehat{\mathbf{g}}_2 \end{bmatrix} \\ &= \left(\mathbf{B}\mathbf{G}_0^{-1} \otimes \mathbf{X}'\mathbf{G}^{-1}\right) \mathbf{V}^{-1} \left(\begin{bmatrix} \mathbf{y}_1 - \mathbf{1}\widehat{\mu}_1 \\ \mathbf{y}_2 - \mathbf{1}\widehat{\mu}_2 \end{bmatrix} \right), \end{aligned} \tag{38}$$

with $\widehat{\boldsymbol{\beta}}_i = \widehat{E}\left(\boldsymbol{\beta}_i|\mathbf{y}_1, \mathbf{y}_2\right)$, $i = 1, 2$. After lengthy algebra and using Henderson (1975), the prediction error variance-covariance matrix of the BLUP of marker effects is given by

$$\begin{aligned} & Var\left(\left[\begin{array}{c} \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{array}\right]\right) \\ = \; & (\mathbf{B} \otimes \mathbf{I}_p) - (\mathbf{B} \otimes \mathbf{X}') \left[\mathbf{I}_{2n} + \mathbf{V}^{-1} - \frac{\mathbf{V}^{-1}\mathbf{1}\mathbf{1}'\mathbf{V}^{-1}}{\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}}\right] (\mathbf{B} \otimes \mathbf{X}')'. \end{aligned} \tag{39}$$

A set of $t - statistics$ can be formed by taking the ratio between the BLUP of a given marker effect as in (38) and the square root of the corresponding diagonal element of (39). The statistic is a crude criterion for association between marker and phenotype as it ignores uncertainty associated with the fact that $\mathbf{B}$ and $\mathbf{R}_0$ are estimated from the data, as opposed to being "true values" required by BLUP theory.

The Bayesian bivariate GBLUP model used standard assumption as in Sorensen and Gianola (2002), i.e., it was a multivariate normal-inverse Wishart hierarchical specification. The only difference with GBLUP is that, in the Bayesian treatment, $\mathbf{G}_0$ and $\mathbf{R}_0$ were treated as unknown parameters, with the uncertainty about their values accounted for.

6.2 Bivariate LASSO

Our MCMC implementation for MBL was applied to markers directly, as opposed to inferring their effects from signal indirectly, as it is done for GBLUP. The model was as in (4) with $T = 2$. Each marker was assigned a conditional bivariate Laplace prior distribution with scale matrix $\boldsymbol{\Sigma}$; in turn, $\boldsymbol{\Sigma}$ was given a two-dimensional inverse Wishart distribution on $\nu_\beta = 4$ degrees of freedom and with scale matrix $\boldsymbol{\Omega}_\beta = \nu_\beta \mathbf{B}/12 = \mathbf{B}/3$. The residual variance-covariance matrix $\mathbf{R}_0$ was assigned the two-dimensional Jeffreys improper prior in (6).

The MCMC scheme employed the scale mixture of normals representation of the bivariate Laplace distribution. First, six independent chains of 1500 iterations each were run. The shrinkage diagnostic metric of Gelman and Rubin (1992) was calculated for μ_1, μ_2, $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$, for the effect of marker 10 on trait 1, and for the effect of marker 200 on trait 2; the `R` package CODA was used for this purpose. Supplementary Figures 4-13 gave no strong evidence of lack of convergence, as indicated by shrinkage factor values close to 1.

Post-burn in samples were collected for an additional 2000 iterations in each chain, so a total of 12,000 samples (without thinning) was used for inference. Supplementary Figures 14 and 15 depict post burn-in trace plots for the elements of $\mathbf{R}_0$ and $\boldsymbol{\Sigma}$, respectively. The six chains "joined" eventually and sample values thereafter fluctuated within what seemed to be stationary distributions. To assess convergence further, a test suggested by Geweke (1992) was applied to the combined 12000 samples from the posterior distributions of $\mu_1, r_{e_{12}}$ (residual correlation

between yields in environments 1 and 2) and $r_{\beta_{12}}$, the correlation between effects of a marker on the two traits. The test compared means of two parts of the combined collection of 12,000 samples at each of 10 segments of the collection: there was no evidence of lack of convergence. In short, the implementation met successfully the convergence tests applied.

Figures 1 and 2 display estimated posterior densities of $r_{e_{12}}$ and $r_{\beta_{12}}$. Mixing for $r_{\beta_{12}}$ was poorer than for $r_{e_{12}}$; the effective number of samples was 220.6 and 979.0, respectively, and Monte Carlo errors were small enough. The residual correlation (posterior mean, 0.17) was positive and different from 0, whereas the $r_{\beta_{12}}$ parameter was estimated at -0.35, also different from zero. However, the posterior densities were not sharp enough for precise inference, probably due to the small sample size ($n = 599$) and low density of the marker panel ($p = 1279$). The quality of these estimates is of subsidiary interest here as our objective was to demonstrate the MBL in a comparison with bibariate BLUP of marker effects.

Location parameters mixed well. For example, the average effective sample size of μ_1 over the 6 chains during burn in was 1499 for a nominal 1500 iterations. For marker 10 effect on trait 1 it was 962 out of 1500, and for marker 200 effect on trait 2 effective size was 1130 out of 1500. These numbers suggest that all 2558 marker effects were estimated with a very small Monte Carlo error in our MBL implementation with 12,000 samples used for inference.

6.3 MBL vs BLUP estimates of marker effects

Figure 3 gives a comparison between bivariate BLUP and posterior mean estimates of effects from MBL. The upper panel shows good alignment between estimates, except at the extremes of the scatter plots. The lower panel depicts that markers with the strongest absolute effects, as estimated by BLUP, had an even stronger effect when estimated under the bivariate BL. Figure 4 presents standardized estimates of each of the 1279 marker effects, by trait. For GBLUP the $t - statistic$ was the estimated marker effect divided by the square root of its prediction error variance; for MBL it was the posterior mean divided by its posterior standard deviation. There is no evidence that any of the markers had an effect differing from 0, corroborating the view that wheat yield is a typical quantitative trait affected by many variants each having a small effects (Singh et al. 1986; Sleper and Poehlman 2006). Using a univariate least-squares, GWAS-type analysis, there were 29 (yield 1) and 56 (yield 2) significant hits after a Bonferroni correction (1279 tests, $\alpha = 0.05$). A comparison between the $t - statistics$ from the GWAS-type analysis with the standardized BLUP and MBL effects is provided in Figure 5. As expected, shrinkage towards null-mean distributions (bivariate Gaussian in BLUP and bivariate Laplace in MBL) made $t - statistics$ much smaller in absolute value than the corresponding ones from GWAS.

Standard GWAS aims to find connections between markers and genomic regions having an effect on a single trait (e.g., Manolio et al. 2009, Visscher et al. 2012; Gianola et al. 2016; Schaid et al. 2018) A search for pleiotropy, on the other hand, focuses on markers having multi-trait

effects. The latter can be viewed as a search for vectors of effects with non-null coordinates that are distant from a $T-$dimensional 0 origin. Mahalanobis squared distances $\left(m_j^2\right)$ away from $(0,0)$ for each the 1279 bivariate vectors of marker effects were calculated for both BLUP and MBL. For BLUP and marker j, the squared distance was computed as $m^2_{Blup,j} = \boldsymbol{\beta}'_{Blup,j}\mathbf{B}^{-1}\boldsymbol{\beta}_{Blup,j}$, and for MBL it was $m^2_{MBL,j} = \boldsymbol{\beta}'_{MBL,j}\left(12\overline{\boldsymbol{\Sigma}}\right)^{-1}\boldsymbol{\beta}_{MBL,j}$ where $\boldsymbol{\beta}_{.,j}$ are effect estimates for marker j and $\overline{\boldsymbol{\Sigma}}$ is the estimated posterior expectation of $\boldsymbol{\Sigma}$. For BLUP, $m^2_{Blup,j}$ had median and maximum values of 0.16 and 2.94, respectively, over markers. For MBL the corresponding values were 0.14 and 3.83. Figure 6 shows that the largest estimated distances were obtained with MBL, supporting the view that the method produces less shrinkage of multiple-trait effect sizes than BLUP. If the 95% percentile of a chi-squared distribution on 2 degrees of freedom (5.99 and 14.4 without and with a Bonferroni correction at $\alpha = 0.05$) is used as "significance threshold", none of the 1279 markers could be claimed to have a bivariate effect on the trait, which is consistent with the $t-statistics$.

7 Predictive comparison between MBL, MTGBLUP and MT-BayesCπ: wheat

Bivariate Bayesian GBLUP and BayesCπ models (Cheng et al. 2018a) were also fitted to the wheat data set. Multiple-trait Bayesian linear models are well known (e.g., Sorensen and Gianola 2002); BayesCπ consisted of a bivariate mixture in which each of the 1279 markers was allowed to fall, *a priori*, into one of four disjoint classes: $(0,0)\,,(0,1)\,,(1,0)\,,(1,1)$, where $(0,0)$ means that a marker has no effect on either trait, $(0,1)$ indicates that a marker affects yield 2 only, and so on. The prior for the four probabilities of membership was a $Dirichlet\,(1,1,1,1)$ distribution. All three methods were run in each of 100 randomly constructed training sets and predictions were formed for lines in corresponding testing sets. Training and testing set sizes had 300 and 299 wheat lines, respectively, in each of the 100 runs. For all methods, the MCMC scheme was a single long chain of 50,000 iterations, with a burn-in period of 1,000 draws. The analyses were run using the JWAS package written in the JULIA language (Cheng et al. 2018b).

Figures 7 and 8 present pairwise plots (bivariate Bayesian GBLUP denoted as RR-BLUP in the plots) of predictive correlations and predictive mean-squared errors, respectively; the plots display less than 100 (X,Y) points because numbers were rounded to two decimal points. There were no appreciable differences in predictive performance between the three methods, supporting the view that cereal grain yield is multi-factorial and that there are none, if any, genomic regions, with large effects. The variability among replications of the training-testing layout is essentially random, reinforcing the notion of the importance of measuring uncertainty of prediction (Gianola et al. 2018). Many studies fail to replicate and often claim differences

between methods based on single realizations of predictive analyses.

8 Predictive comparison between MBL vs MTGBLUP and MBL vs single trait Bayesian LASSO: *Pinus*

Figures 9 and 10 present scatter-plots of the predictive performance (mean squared error and correlation, respectively) of the bivariate Bayesian LASSO and bivariate Bayesian GBLUP (MTGBLUP, denoted as RR-BLUP in the plots) in the 100 testing sets. There were no obvious differences in mean-squared error for either rust bin or gall volume although, for the latter trait, a slight superiority of MBL was noted (Figure 9); the plot contains distinct 12 points because the overlap in numerical values produced "clusters" of points. On the other hand, there was a decisive superiority (Figure 10) of MBL over MTGBLUP in predictive correlation.

Figure 11 contrasts the predictive performance of the bivariate Bayesian LASSO over the single trait Bayesian LASSO for gall volume. The two trait analysis tended to produce larger predictive correlations and smaller mean-squared errors, illustrating an instances in which a multiple-trait specification clearly constitutes a better prediction machine.

9 Conclusion

Our study is possibly the first report in the quantitative genetics literature of a multiple-trait Bayesian LASSO (MBL), inspired by the BL of Park and Casella (2008). MBL assumes that vectors of marker effects on T traits follow a null-mean multivariate Laplace distribution, *a priori*. This distribution has a sharp peak at the origin and reduces to the double exponential prior of the BL when applied to a single trait. The implementation of MBL requires Markov chain Monte Carlo sampling and a relative simple Metropolis-Hastings algorithm based on a scaled mixture of normals representation (Gómez-Sánchez-Manzano et al. 2008) was presented. The algorithm was tested thoroughly with a wheat data set and found to mix well, with no evidence of lack of convergence to the posterior distribution and with a small Monte Carlo error.

A question that arises often in practice, is the extent to which a multiple-trait method will produce a better performance than a single-trait specification. If the parameters of the model (assuming it holds) representing the inter-trait distribution are either known or well estimated, one should expect more power for QTL detection and a better predictive performance for the multivariate specification. In our study we found that MBL outperformed the single trait in terms of delivering a better predictive performance for gall volume but not for rust bin in *Pinus.* On the other hand, a multiple-trait analysis is more complex and requires more assumptions, so it may be less robust than a single trait procedure and fail to deliver according

to expectation in real-life circumstances. It is risky to make sweeping statements arguing in favor of a specific treatment of data as outcomes are heavily dependent on the biological architecture of the traits considered, and on the data structure as well. The picture emerging from two decades of experience in genome-enabled prediction in the fields of animal and plant breeding is that is largely futile to categorize methods in terms of expected predictive performance using broad criteria, in view of the large variability of performance with respect to data structure for any given prediction machine (Morota and Gianola 2014; Gianola and Rosa 2015; Momen et al. 2018; Montesinos-López et al. 2019 a,b,c,d; Azodi et al. 2019).

MBL is expected to shrink more strongly towards zero vectors of markers with small effects in their coordinates, thus producing differential shrinkage and preserving the *modus operandi* of BL. Mimicking the single-trait argument in Tibshirani (1996) which shows equivalence between LASSO and a posterior mode, the representation in Appendix E illustrates that the degree of shrinkage of the vectorial effects of a marker (j, say) on a set of traits is inversely proportional to the quadratic form $\boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j$. Thus, multivariate Bayesian pseudo-sparsity is induced by MBL to an extent depending on the heterogeneity of $\boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j$ over markers. We note, in passing, that the term $\sum_{j=1}^{p}\sqrt{\boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j}$ given in (66) of Appendix E is the counterpart of $\sum_{g=1}^{G}\sqrt{\sum_{t=1}^{T}\beta_{gt}^2}$, part of the "group-penalty" in Li et al. (2015), where g is some meaningful group of markers arrived at, say, on the basis of biological considerations, and β_{gt}^2 is the group regression coefficient for trait t. The latter penalty assigns the same weight to these regressions over trains, contrary to MBL where weights and co-weights are driven by $\boldsymbol{\Sigma}^{-1}$. The BL or MBL can be adapted to situations where a group structure may be needed via hierarchical modeling; this fairly straightforward issue is outside of the scope of the paper but may pursued in future extensions of MBL. Actually, Liquet et al. (2017) described a Bayesian multiple-trait analysis where a LASSO-type penalty is assigned to group effects and a spike-slab prior induces additional Bayesian sparsity at the level of individual regression coefficients. The authors did not address the predictive ability of their method so it would be interesting to compare it against MBL and the multiple-trait mixture model of Cheng et al. (2018). We plan to carry out this comparison in collaboration with CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo, México) using a large number of data sets in various cereal crops.

Knowledge of the genetic basis of complex traits is limited and not vast enough to enable formulation of *a priori* prescriptions for any specific trait or situation. The number, location and effects of causal variants, the linkage disequilibrium structure between such variants and markers, and the mode of gene action of QTL are largely unknown, this holding for all species of domesticated plants and animals and for most common diseases in humans. Theoretically, MBL is expected to perform better than multiple-trait BLUP whenever appreciable heterogeneity exists over the effects of the markers in the panel employed, while behaving as multiple-trait

GBLUP when all markers have tiny and similar effects. This consideration follows directly from the structure of the method, and computer simulations could be easily tailored to create scenarios where MBL has a better or a worse performance simply by design but without necessarily being relevant to a real-life inferential or predictive problem.

The expectation stated above was verified empirically: markers with stronger (positive or negative) effects on the wheat yields examined had larger Mahalanobis distances away from zero than markers with small effects. Further, markers with short distances in GBLUP had even shorter distances under MBL. None of the two methods was able to detect variants having a strong effect on wheat yield, contrary to least-squares GWAS. However, outcomes from GWAS are not strictly comparable with those from shrinkage-based procedures. In single-marker least-squares the estimator is potentially biased because other genomic regions are ignored in the model; further, short and long range linkage disequilibria create statistical ambiguity (Gianola et al. 2016). In WGR, on the other hand, regressions are akin to partial derivatives, i.e., the coefficient gives the net effect of the marker given that the other markers are fitted; typically, regressions become smaller as p is increased at a fixed n.

In plant and animal breeding, a focal point is the evaluation of genetic merit of candidates for artificial selection, and the prediction of expected performance in either collateral relatives or in descendants. Under the assumptions of additive inheritance, genome-enabled prediction (Meuwissen et al. 2001) produces estimates of marked additive genomic value, $\mathbf{g}$, or signal as referred to in our paper. In MBL, $\mathbf{g}$ and marker effects can be inferred from their posterior mean or from a modal approximation (MAP-MBL) that does not involve MCMC which is described in Appendix E. A rough comparison between GBLUP, MBL and MAP-MBL was carried out with the wheat data. For the latter, we used $\boldsymbol{\Sigma} = \mathbf{G}_0/(12p)$, and starting values for the iteration were calculated using BLUP estimates of marker effects. MAP with $T = 2$ were iterated for 500 rounds. Supplementary Figure S16 shows that, at iteration 500, the metric used for monitoring convergence had stabilized at the third decimal place, but iteration could have stopped after 200 rounds, for our purposes. Supplementary Figure S17 presents a scatter plot of the 2558 (bivariate) marker effect solutions at iterations 1 and 500 against the corresponding BLUP or MBL posterior mean estimates. Clearly, MAP approach gave markedly different results, producing a stronger shrinkage to 0 of small-effect markers an, thus, an effectively more sparse model. Supplementary Figure S18 gives a comparison of the fitted genomic values, i.e., $\mathbf{g}_1 = \mathbf{X}\boldsymbol{\beta}_1^*$ and $\mathbf{g}_2 = \mathbf{X}\boldsymbol{\beta}_2^*$ for the two traits. GBLUP and MBL estimates were closely aligned and fitted the data in a similar manner. On the other hand, MAP-MBL gave a larger mean-squared error of fit and a smaller correlation between fitted and observed phenotypes, possibly because of the larger effective sparsity of MAP-MBL. A worse fit to the data does not necessarily imply a poorer predictive ability. A thorough comparison of predictive ability between MBL and MAP-MBL will be carried out in future research.

Our predictive comparison in wheat involved three bivariate models: GBLUP, MBL and BayesCπ, which employs Bayesian model averaging. A training-testing validation replicated 100 times at random indicated no differences among methods. However, it was found that MBL was better than MT Bayesian BLUP for the two pine tree traits. After almost two decades of genome-enabled prediction it is now clear that no universally best prediction machine exists (Gianola et al. 2011; Heslot 2012; de los Campos et al. 2013; Momen et al. 2018; Bellot et al. 2018; Montesinos-López et al. 2018a, b, c, d) even when non-parametric or deep learning techniques are brought into the comparisons.

As far as we know, our paper represents the first report in the quantitative genetics literature of a multiple-trait LASSO, implemented in a Bayesian or empirical Bayes (Appendix E) manner. MBL adds to the armamentarium of genome-enabled prediction and expands the family of members of the Bayesian alphabet (Gianola et al. 2009; Habier et al. 2011; Gianola 2013). Further, it has been implemented in the publicly available JWAS software (Cheng et al. 2018b). We take the view that every prediction problem is unique and that no claims about the superiority of a specific procedure over others should be made without qualifications. For instance, MBL could perform worse or better than here when applied to other species, traits, or when confronted with different data structures. Most quantitative and disease traits are truly complex and it is dangerous to offer simplistic solutions or predictive panaceas (Goddard et al. 2019).

10 Acknowledgments

Professor Chris-Carolin is thanked for useful discussions and comments. J. M. Marín provided technical information on the MLAP distribution. Two anonymous reviewers are thanked for helpful suggestions. Financial support to DG was provided by the Deutsche Forschungsgemeinschaft (grant No. SCHO 690/4-1 to CCS) and by the J. Lush Endowment as Visiting Professor at Iowa State University in 2018. The University of Wisconsin-Madison, USA, the Technical University of Munich (TUM) Institute for Advanced Study, TUM School of Life Sciences, the Institut Pasteur de Montevideo, Uruguay, are thanked for providing office space, library and computing resources and logistic support.

11 Legend for Figures

Figure 1. Bivariate Bayesian LASSO: trace plot and posterior density of residual correlation.

Figure 2. Bivariate Bayesian LASSO: trace plot and posterior density of correlation between marker effects.

Figure 3. Bivariate GBLUP versus bivariate Bayesian LASSO (posterior mean) estimates of marker effects on wheat grain yield.

Figure 4. t-statistics for marker effects on wheat grain yield: GBLUP versus bivariate Bayesian LASSO (MBL)

Figure 5. t-statistics for marker effects on wheat grain yield: ordinary leas-squares (OLS) versus bivariate Bayesian LASSO (MBL) and bivariate GBLUP.

Figure 6. Mahalanobis squared distance (M) away from (0,0) for bivariate effects on grain yield of 1279 markers: GBLUP versus bivariate Bayesian LASSO (BLASSO)

Figure 7. Predictive correlations for wheat grain yield: bivariate Bayesian LASSO (Bayes L) versus bivariate GBLUP (RR-BLUP).

Figure 8. Predictive mean-squared error for grain yield: bivariate Bayesian LASSO (Bayes L), bivariate GBLUP (RR-BLUP) and bivariate Bayes Cπ.

Figure 9. Mean-squared error of prediction for rust bin and gall volume in pine trees: bivariate Bayesian LASSO (Bayes L) versus bivariate Bayesian GBLUP (RR-BLUP).

Figure 10. Predictive correlation for rust bin and gall volume in pine trees: bivariate Bayesian LASSO (Bayes L) versus bivariate Bayesian GBLUP (RR-BLUP).

Figure 11. Predictive mean squared error and correlation for gall volue in pine trees: bivariate Bayesian LASSO (MTBayesL) versus univariate Bayesian LASSO (STBayesL)

12 Literature Cited

Azodi, C., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., Shiu, S-H., 2019, Benchmarking parametric and machine learning models for genomic prediction of complex traits. Genes, Genomes, Genetics 9: 3691-3702.

Calus, M. P. L., Veerkamp, R. F., 2011, Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution. doi.org/10.1186/1297-9686-43-26.

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., 2018a, Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics 209: 89-103.

Cheng, H., Fernando, R., Garrick, D., 2018b, Julia implementation of whole-genome analyses Software. Proceedings of the World Congress on Genetics Applied to Livestock Production. http://www.wcgalp.org/proceedings/2018/jwas-julia-implementation-whole-genome-analyses-software

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., et al. 2010, Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724.

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

de los Campos G., Gianola, D., Allison, D. A. B., 2011, Predicting genetic predisposition in humans: the promise of whole-genome markers. Nature Reviews Genetics 11: 880-886.

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

Falconer, D. S., 1952, The problem of environment and selection. American Naturalist 86: 293-298.

Fernando, R., Toosi, A., Wolc, A., Garrick, D., Dekkers, J., 2017, Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. Journal of Agricultural, Biological and Environmental Statistics 22: 172–193.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M, Janss, L. L.,Vermeulen, S. H., 2014. A comparison of multivariate genome-wide association methods. PLoS One.

doi.org/10.1371/journal.pone.0095923

Gao, H., Madsen, P., Pösö, J., Aamand, G. P., Lidauer, M, Jensen, J., 2018. Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population. Journal of Dairy Science 101: 11159-11164.

Gelfand, A. E., Dey, D. K., Chang, H., 1992, Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics* 4, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 147-167. Oxford University Press.

Gelman, A., Rubin, D. B., 1992, Inference from iterative simulation using multiple sequences (with discussion). Statistical Science 7: 457-511.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2014, *Bayesian Data Analysis.* 3rd Edition. Chapman and Hall/CRC Press. Boca Raton.

Geweke, J.,1992, Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (ed. Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. Clarendon Press, Oxford.

Gianola, D., 2013, Priors in whole genome regression: the Bayesian alphabet returns. Genetics 194: 573-596.

Gianola, D., Pérez-Enciso, M., Toro, M. A., 2003, On marker-assisted prediction of genetic value: beyond the ridge. Genetics 63: 347-65.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., Fernando, R. L., 2009, Additive genetic variability and the Bayesian alphabet. Genetics 187: 347–63.

Gianola, D., Okut, H., Weigel, K. A., Rosa, G. J. M., 2011, Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics. doi: 10.1186/1471-2156-12-87.

Gianola, D., de los Campos, G., Toro, M. A., Naya, H., Schön, C-C, Sorensen, D., 2015, Do molecular markers inform about pleiotropy? Genetics 201: 23-99.

Gianola, D., Rosa, G. J. M., 2015, One hundred years of statistical developments in animal breeding. Annual Reviews of Animal Biosciences 3:19-56.

Gianola, D., Fariello, M. I., Naya, H., Schön, C-C, 2016, Genome-wide association studies with a genomic relationship matrix: a case study with wheat and *Arabidopsis*. G3: Genes, Genomes. Genetics doi: 10.1534/g3.116.034256.

Gianola, D., Cecchinato, A., Naya, H., Schön, C-C, 2018, Prediction of complex traits: robust alternatives to best linear unbiased prediction. Frontiers in Genetics doi.org/10.3389/fgene.2018.00195.

Goddard, M. E., 2009, Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 2452-57.

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., Hayes, B. J., 2016, Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proceedings of the Royal Society B. doi.org/10.1098/rspb.2016.0569.

Goddard M. E., Meuwissen, T. H. E., Daetwyler, H. D., 2019, Prediction of phenotype from DNA variants. In *Handbook of Statistical Genomics,* Eds. Balding, D., Moltke, I., Marioni, J. Wiley, Hoboken, USA and Chichester, UK.

Gómez, E ., Gómez-Villegas, M. A., Marín, J. M., 2007, A multivariate generalization of the power exponential family of distributions. Communications in Statistics - Theory and Methods 27: 589-600.

Gómez-Sánchez-Manzano, E. , Gómez-Villegas, M. A., Gómez, E., Marín, J. M., 2008, Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. Communications in Statistics - Theory and Methods 37: 972-985.

Habier, D., Fernando, R. L., Kizilkaya, K., Garrick, D. J., 2011, Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. doi.org/10.1186/1471-2105-12-18.

Hazel, L. N., 1943, The genetic basis for constructing selection indexes. Genetics 28: 476–490.

Henderson, C. R., 1975, Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–449.

Henderson, C. R., 1977, Best linear unbiased prediction of breeding values not in the model for records. Journal of Dairy Science 60: 783–787.

Henderson, C. R., Quaas, R. L., 1976 Multiple trait evaluation using relatives' records. Journal of Animal Science 43: 1188–1197.

Heslot, N., Yang, H. P., Sorrells, M. E., Jannink, J. L., 2012, Genomic selection in plant breeding: a comparison of models. Crop Science 52: 146-160.

Isik, F., Holland, J., Maltecca, C., 2017, *Genetic Data Analysis for Plant and Animal Breeding.* Springer, Cham.

Jia, Y., Jannink, J. L., 2012, Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192: 1513-1522.

Lande, R., Thompson, R., 1990, Efficiency of marker-assisted selection in the improvement

of quantitative traits. Genetics 124: 743-756.

Lee, S. H., Wray, N. R., Goddard, M. E., Visscher, P. M., 2011, Estimating missing heritability for disease from genome-wide association studies. American Journal of Human Genetics 88: 294-305.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011, Improved LASSO for genomic selection. Genetics Research 93: 77-87.

Lehermeier C., Wimmer, V., Albrecht, T., Auinger, H., Gianola, D., Schön, C-C, Schmid, V. J., 2013, Sensitivity to prior specification in Bayesian genome-based prediction models. Statistical Applications in Genetics and Molecular Biology 12: 1–17.

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

Li, Z., Möttoïen, J., Sillanpää, M. J., 2015, A robust multiple-trait locus method for quantitative trait locus analysis of non-normally distributed multiple traits. Heredity 115: 556-564.

Liquet, B., Mengersen, K. Pettitt, A. N., Sutton, M., 2017, Bayesian variable selection regression of multivariate responses for group data. Bayesian Analysis 12: 1039-1067.

Long, N., Gianola. D., Rosa, G. J. M., Weigel, K. A., 2011, Marker-assisted prediction of non-additive genetic values. Genetica 139: 843-854.

López de Maturana, E., Chanok, S. J., Picornell, A. C., Rothman, N., Herranz, J., Calle, M. L., Garcia-Closas, M., Marenne, G., Brand, A., Tardon, A., Carrato, A., Silverman, D. T., Kogevinas, M., Gianola, D., Real, F.X., Malats, N., 2014, Whole genome prediction of bladder cancer risk with the Bayesian LASSO. Genetic Epidemiology. DOI 10.1002/gepi.21809.

Makowsky R., Pajewski, N. M., Klimentidis, Y. C., Vázquez, A. I., Duarte, C. W., Allison, D. A. B., de los Campos, G., 2011, Beyond missing heritability: prediction of complex traits. PLoS Genetics doi:10.1371/journal.pgen.1002051.

Manolio, T.A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., et al., 2009, Finding the missing heritability of complex diseases. Nature 461: 747–753.

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

Momen, M., Ayatollahi-Mehrgardi, A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., Rosa, G. J. M., Gianola, D., 2018, Predictive ability of genome-assisted statistical models under various forms of gene action. Scientific reports 8: 1-11.

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., et al., 2019a, New deep learning genomic-based prediction model for multiple traits with binary, ordinal, and continuous phenotypes. G3: Genes, Genomes, Genetics 9: 1545-1556.

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, C. M., et al., 2019b, A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. G3:

Genes, Genomes, Genetics 9: 601-618.

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., Hernández-Suárez, C. M., 2019c, Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3: Genes, Genomes, Genetics 8: 3813-3828.

Montesinos-López, O. A., Martín-Vallejo, J., Crossa, J., Gianola, D., Hernández-Suárez, 2019d, Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3: Genes, Genomes, Genetics 8: 3829-3840.

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. M., Visscher, P. M., 2015, Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture mode. PLoS Genetics. doi.org/10.1371/journal.pgen.1004969.

Park, T., Casella G., 2008, The Bayesian LASSO. Journal of the American Statistical Association 103: 681-686.

Pérez, P., de los Campos, G., 2014, Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483-495.

Samorodnitsky, G., Taqqu, M. S., 2000. *Stable non-Gaussian random processes*. Chapman-Hall/CRC Boca Raton.

Schaid, D. J., Wenan, C., Larson, N. B., 2018, From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19: 491–504.

Singh G., Bhullar, G. S., Gill, K. S., 1986, Genetic control of grain yield and its related traits in bread wheat. Theoretical & Applied Genetics 72: 536-40.

Sleper, D. A., Poehlman M. (2006) *Breeding Field Crops*. 5th Edition. Wiley-Blackwell.

Smith, H. F., 1936, A discriminant function for plant selection. Annals of Eugenics. 7: 240–250.

Sorensen, D., Gianola, D., 2002, *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer, New York.

Strandén, I., Garrick, D. J., 2009, Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of Dairy Science 92: 2971-2975.

Tibshirani, R. 1996, Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society B 58: 267-288.

Van Raden P. M., 2007, Genomic measures of relationship and inbreeding. Interbull Bulletin 37: 33–36.

Van Raden P. M., 2008, Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414–4423.

Van Tassell, C. P., Van Vleck, L. D., 1996, Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. Journal of Animal Science 74: 2586-2597.

Visscher, P. M., Brown, M. A., McCarthy, M. I., Yang, J., 2012, Five years of GWAS

781 discovery. American Journal of Human Genetics 13: 7-24.

782 Visscher, P. M., Wray, N. R., Zhang, Q., et al., 2017, 10 Years of GWAS Discovery: Biology,
783 Function, and Translation. American Journal of Human Genetics 101: 5-22.

784 Walsh, B., Lynch, M., 2018, *Evolution and selection of quantitative traits*. Oxford University
785 Press, New York.

786 Yang J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden,
787 P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., Visscher, P. M., 2010,
788 Common SNP's explain a large proportion of the heritability for human height. Nature Genetics
789 42: 565–569.

790 Yi, N., Xu, S., 2008, Bayesian LASSO for quantitative trait loci mapping. Genetics 179:
791 1045-1055.

792 Yuan, M., Lin, Y., 2006, Model selection and estimation in regression with grouped variables.
793 Journal of the Royal Statistical Society B 68: 49-67.

794 Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007, Dimension reduction and coefficient estima-
795 tion in multivariate linear regression. Journal of the Royal Statistical Society B 69: 329-346.

13 Appendices

13.1 Appendix A: Excursus on the MLAP distribution

13.1.1 Three bivariate Laplace distributions

For illustration, consider three bivariate Laplace distributions, all having null means but distinct scale matrices, as follows:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}; \ \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}. \tag{40}$$

Using (7), the density under $\boldsymbol{\Sigma}_1$ is

$$p\left(\beta_1, \beta_2 | \boldsymbol{\Sigma}_1\right) = \frac{1}{8\pi} \exp\left(-\frac{1}{2}\sqrt{\beta_1^2 + \beta_2^2}\right). \tag{41}$$

The covariance matrix here, $\mathbf{B}_1 = 12\boldsymbol{\Sigma}_1$, is diagonal, so the random variables are uncorrelated but not independent since (41) cannot be written as the product of two marginal densities. Under $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$, the densities are

$$p\left(\beta_1, \beta_2 | \boldsymbol{\Sigma}_2\right) = \frac{5\sqrt{6}}{96\pi} \exp\left(-\frac{1}{2}\sqrt{\frac{25}{24}\beta_1^2 - \frac{5}{12}\beta_1\beta_2 + \frac{25}{24}\beta_2^2}\right), \tag{42}$$

and

$$p\left(\beta_1, \beta_2 | \boldsymbol{\Sigma}_3\right) = \frac{5}{24\pi} \exp\left(-\frac{1}{2}\sqrt{\left(\frac{25}{9}\beta_1^2 + \frac{40}{9}\beta_1\beta_2 + \frac{25}{9}\beta_2^2\right)}\right), \tag{43}$$

Five bivariate Laplace densities are shown in Supplementary Figure 1. (a) gives the density of the distribution of the two uncorrelated bivariate Laplace random variables ($\boldsymbol{\Sigma}_1$), and (b) and (c) show the positively (i.e., with $\boldsymbol{\Sigma}_2$) and negatively (with $\boldsymbol{\Sigma}_3$) correlated situations, respectively. These three densities have a sharp mode at $\beta_1 = \beta_2 = 0$ indicating that a bivariate Laplace prior would strongly shrink vectors to the $(0, 0)$ point, acting similarly to the DE prior in the univariate Bayesian LASSO. (d) and (e) displays bivariate Laplace densities of distributions with non-null means.

13.1.2 Conditional distributions

Dropping subscript j denoting a specific marker, partition the $T \times 1$ vector of effects into $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ where the sub-vectors have orders T_1 and T_2, respectively; recall that T is the number of traits. Correspondingly, put

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{44}$$

According to J. M. Marín (personal communication), the conditional distribution $[\boldsymbol{\beta}_2|\boldsymbol{\beta}_1]$ has density

$$p\left(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1,\boldsymbol{\mu}_{2|1},\boldsymbol{\Sigma}_{2|1}\right) = \frac{\Gamma\left(\frac{T-T_1}{2}\right)}{|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}\pi^{\frac{T-T_1}{2}}\int_0^\infty t^{\frac{T-T_1}{2}-1}\exp\left\{-\frac{1}{2}\sqrt{t+q_1}\right\}dt} \times \exp\left\{-\frac{1}{2}\left[\sqrt{\left(\boldsymbol{\beta}_2-\boldsymbol{\mu}_{2|1}\right)'\boldsymbol{\Sigma}_{2|1}^{-1}\left(\boldsymbol{\beta}_2-\boldsymbol{\mu}_{2|1}\right)+q_1}\right]\right\}, \quad (45)$$

where $\boldsymbol{\mu}_{2|1}=\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1$, $\boldsymbol{\Sigma}_{2|1}=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $q_1=\boldsymbol{\beta}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1$. Similar to multivariate normal distribution, the conditional expectation is linear on the conditioning variable and $\boldsymbol{\Sigma}_{2|1}$ does not involve $\boldsymbol{\beta}$.

13.1.3 Simulation of a multivariate Laplace distribution

Gómez et al. (2007) showed that S independent draws from a MLAP distribution with a null mean vector can be made as

$$\boldsymbol{\beta}_i = r_i\mathbf{C}'\mathbf{u}_i;\ i=1,2,...,S, \quad (46)$$

where $\mathbf{C}'$ results from the Cholesky decomposition $\boldsymbol{\Sigma}=\mathbf{C}'\mathbf{C}$, $\mathbf{u}$ is a $T\times 1$ vector uniformly distributed on a T-dimensional unit sphere and r is a realization of a Gamma distribution with shape parameter T and scale 2. Vector $\mathbf{u}$ can be simulated by effecting T independent draws $(x_i;\ i=1,2,...,T)$ from a $N(0,1)$ distribution, and then forming the t^{th} element of $\mathbf{u}$ as $u_t=x_t/\sqrt{\sum_{t=1}^{T}x_t^2}$, $t=1,2,...,T$.

Marginal distributions for the three bivariate Laplace distributions with scale matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ given above were estimated by sampling $S=300,000$ independent realizations; (46) was employed. Using the samples, zero-mean DE and normal distributions with the same variances were fitted, and the resulting densities were compared with the estimated densities based on the draws. As shown in Supplementary Figure 2, a normal distribution provided a poor approximation to the marginals from the three bivariate Laplace cases, and the sharp peak at 0, characteristic of a DE density, was not matched by such marginals. This is a corroboration of theoretical results in Gómez et al. (2007): marginals from MLAP distributions are elliptically contoured and not DE.

13.2 Appendix B: Mean vector of location parameters given ELSE

Consider (19). For $T = 3$, let

$$\mathbf{R}_0^{-1} = \begin{bmatrix} r^{11} & r^{12} & r^{13} \\ r^{21} & r^{22} & r^{23} \\ r^{31} & r^{32} & r^{33} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} & \Sigma^{13} \\ \Sigma^{21} & \Sigma^{22} & \Sigma^{23} \\ \Sigma^{31} & \Sigma^{32} & \Sigma^{33} \end{bmatrix} \tag{47}$$

Expansion of the Kronecker products in (19) produces the system

$$\begin{bmatrix} r^{11}N & r^{12}N & r^{13}N & r^{11}\mathbf{1}_N'\mathbf{X} & r^{12}\mathbf{1}_N'\mathbf{X} & r^{13}\mathbf{1}_N'\mathbf{X} \\ r^{21}N & r^{22}N & r^{23}N & r^{21}\mathbf{1}_N'\mathbf{X} & r^{22}\mathbf{1}_N'\mathbf{X} & r^{23}\mathbf{1}_N'\mathbf{X} \\ r^{31}N & r^{32}N & r^{33}N & r^{31}\mathbf{1}_N'\mathbf{X} & r^{32}\mathbf{1}_N'\mathbf{X} & r^{33}\mathbf{1}_N'\mathbf{X} \\ r^{11}\mathbf{X}'\mathbf{1}_N & r^{12}\mathbf{X}'\mathbf{1}_N & r^{13}\mathbf{X}'\mathbf{1}_N & r^{11}\mathbf{X}'\mathbf{X}+\Sigma^{11}\mathbf{D}^{-1} & r^{12}\mathbf{X}'\mathbf{X}+\Sigma^{12}\mathbf{D}^{-1} & r^{13}\mathbf{X}'\mathbf{X}+\Sigma^{13}\mathbf{D}^{-1} \\ r^{21}\mathbf{X}'\mathbf{1}_N & r^{22}\mathbf{X}'\mathbf{1}_N & r^{23}\mathbf{X}'\mathbf{1}_N & r^{21}\mathbf{X}'\mathbf{X}+\Sigma^{21}\mathbf{D}^{-1} & r^{22}\mathbf{X}'\mathbf{X}+\Sigma^{22}\mathbf{D}^{-1} & r^{23}\mathbf{X}'\mathbf{X}+\Sigma^{23}\mathbf{D}^{-1} \\ r^{31}\mathbf{X}'\mathbf{1}_N & r^{32}\mathbf{X}'\mathbf{1}_N & r^{33}\mathbf{X}'\mathbf{1}_N & r^{31}\mathbf{X}'\mathbf{X}+\Sigma^{31}\mathbf{D}^{-1} & r^{32}\mathbf{X}'\mathbf{X}+\Sigma^{32}\mathbf{D}^{-1} & r^{33}\mathbf{X}'\mathbf{X}+\Sigma^{33}\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \overline{\mu}_1 \\ \overline{\mu}_2 \\ \overline{\mu}_3 \\ \overline{\boldsymbol{\beta}}_1^* \\ \overline{\boldsymbol{\beta}}_2^* \\ \overline{\boldsymbol{\beta}}_3^* \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{1}_N'(r^{11}\mathbf{y}_1^* + r^{12}\mathbf{y}_2^* + r^{13}\mathbf{y}_3^*) \\ \mathbf{1}_N'(r^{21}\mathbf{y}_1^* + r^{22}\mathbf{y}_2^* + r^{23}\mathbf{y}_3^*) \\ \mathbf{1}_N'(r^{31}\mathbf{y}_1^* + r^{32}\mathbf{y}_2^* + r^{33}\mathbf{y}_3^*) \\ \mathbf{X}'(r^{11}\mathbf{y}_1^* + r^{12}\mathbf{y}_2^* + r^{13}\mathbf{y}_3^*) \\ \mathbf{X}'(r^{21}\mathbf{y}_1^* + r^{22}\mathbf{y}_2^* + r^{23}\mathbf{y}_3^*) \\ \mathbf{X}'(r^{31}\mathbf{y}_1^* + r^{32}\mathbf{y}_2^* + r^{33}\mathbf{y}_3^*) \end{bmatrix}. \tag{48}$$

Observe how phenotypes for all traits contribute to the solutions of trait-specific effects.

13.3 Appendix C: Sampling from the conditional posterior distribution of v_j^2

Consider (29). Let $Q = \boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j$ take values $\frac{1}{2}, 1, 4$, and 10, say. Numerical integration of (29) between 0 and 1000 produces 3.5203, 3.040 7, 1.844 3, 1.031 4 as reciprocal of the resulting integration constants, with the normalized densities shown in Supplementary Figure 3. The distributions are skewed, and as Q increases the density becomes flatter.

Let $S_{\frac{1}{2}}(y;\sigma)$ be the Lévy density of a positive random variable Y having a positive stable distribution with parameter σ (Samorodnitsky and Taqqu 2000). From Gómez et al. (2007) and

Gómez-Sánchez-Manzano et al. (2008) the Lévy density is

$$S_{\frac{1}{2}}(y;\sigma) = \frac{\left(\frac{\sigma}{4}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} y^{-\frac{1}{2}-1} \exp\left(-\frac{\sigma}{4y}\right), \tag{49}$$

which is that of an inverse Gamma (IG) distribution with parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{\sigma}{4}$. Consider now the transformation (Gómez et al. 2007) $w_j = 2/v_j^2$ so using (29)

$$\begin{aligned} p(w_j|ELSE) &\propto \left(\frac{2}{w_j}\right)^{-\frac{1}{2}} \exp\left[-\frac{Q_j + \frac{1}{w_j^2}}{\frac{4}{w_j}}\right] \frac{2}{w_j^2} \\ &\propto w_j^{-\frac{1}{2}-1} \exp\left[-\frac{w_j Q_j}{4}\right] \exp\left[-\frac{1}{4w_j}\right]. \end{aligned} \tag{50}$$

Consider a Metropolis-Hastings ratio R using (49) with $\sigma = 1$ as proposal distribution, and (18) let y_j be a proposed value and w_j be a member of the target distribution. The ratio is then

$$\begin{aligned} R &= \frac{y_j^{-\frac{1}{2}-1} \exp\left[-\frac{y_j Q_j}{4}\right] \exp\left[-\frac{1}{4y_j}\right]}{w_j^{-\frac{1}{2}-1} \exp\left[-\frac{w_j Q_j}{4}\right] \exp\left[-\frac{1}{4w_j}\right]} \times \frac{w_j^{-\frac{1}{2}-1} \exp\left(-\frac{1}{4w_j}\right)}{y^{-\frac{1}{2}-1} \exp\left(-\frac{1}{4y}\right)} \\ &= \exp\left[\frac{Q_j}{4}(w_j - y_j)\right]; \; j = 1, 2, ..., p. \end{aligned} \tag{51}$$

Hence if a proposal y_j is drawn from $IG(\alpha = \frac{1}{2}, \beta = \frac{1}{4})$, it can be accepted as belonging to the conditional posterior distribution of w_j, with probability equal to R above. If accepted, a "new" $v_j^2 = 2/w_j$ is a member of $p\left(v_j^2|ELSE\right)$ with probability R as well; otherwise stay with the current v_j^2.

13.4 Appendix D: Alternative algorithm for indirect sampling of marker effects

An alternative sampling scheme that uses an equivalent formulation of the model is presented; a two-trait ($T = 2$) situation is employed for ease of presentation. Let $\mathbf{g}_1 = \mathbf{X}\boldsymbol{\beta}_1^*$ and $\mathbf{g}_2 = \mathbf{X}\boldsymbol{\beta}_2^*$ be the genomic values of the N individuals for each of the traits. A model could be

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \end{pmatrix} + \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix}, \tag{52}$$

where residuals are as before. In a standard genomic best linear unbiased prediction (GBLUP, Van Raden 2008) setting, it is assumed that

$$\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} |\mathbf{G}, \mathbf{G}_0 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{G}_0 \otimes \mathbf{G}, \right) \tag{53}$$

where $\mathbf{G}$ is an $N \times N$ marker-based matrix of "genomic relationships", and

$$\mathbf{G}_0 = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} \\ \sigma_{g_{12}} & \sigma^2_{g_2} \end{bmatrix} \tag{54}$$

is a matrix containing the trait-specific genomic variances and their covariances. Specifically, from the definition of $\mathbf{g}_1$ and $\mathbf{g}_2$, and assuming that $\boldsymbol{\beta}^*_t|\sigma^2_{\beta_t} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_{\beta_t}\right)$ $(t = 1, 2)$

$$Var\left(\mathbf{g}_t|\mathbf{X},\sigma^2_{\beta t}\right) = \frac{1}{p}\mathbf{X}\mathbf{X}'\left(p\sigma^2_{\beta_t}\right) = \mathbf{G}\sigma^2_{g_t};\ t = 1, 2, \tag{55}$$

for $\mathbf{G} = \mathbf{X}\mathbf{X}'/p$ and $\sigma^2_{g_i} = p\sigma^2_{\beta_i}$. Similarly, $Cov\left(\mathbf{g}_1, \mathbf{g}'_2|\mathbf{X},\sigma_{\beta_{12}}\right) = \mathbf{G}\sigma_{g_{12}}$, where $\sigma_{g_{12}} = p\sigma_{\beta_{12}}$ and $\sigma_{\beta_{12}}$ is the covariance between marker effects on traits 1 and 2. Let $\mathbf{B} = \left\{\sigma_{\beta_{tt'}}\right\}$ be the 2×2 variance-covariance matrix of marker effects

For a bivariate Bayesian LASSO model, conditionally on the $p \times 1$ vector $\mathbf{v}^2$, one has

$$\begin{bmatrix} \boldsymbol{\beta}^*_1 \\ \boldsymbol{\beta}^*_2 \end{bmatrix} |\boldsymbol{\Sigma}, \mathbf{v}^2 \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{\Sigma} \otimes \mathbf{D}, \right). \tag{56}$$

Hence

$$\begin{aligned} & \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} |\boldsymbol{\Sigma}, \mathbf{v}^2 \\ \sim\ & N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{X} \end{bmatrix} (\boldsymbol{\Sigma} \otimes \mathbf{D}) \begin{bmatrix} \mathbf{X}' & \mathbf{0} \\ \mathbf{0} & \mathbf{X}' \end{bmatrix} \right) \\ =\ & N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{\Sigma} \otimes \mathbf{X}\mathbf{D}\mathbf{X}' \right). \end{aligned} \tag{57}$$

Let $\mathbf{C}_{Cond} = \boldsymbol{\Sigma} \otimes \mathbf{X}\mathbf{D}\mathbf{X}'$. Further,

$$E\left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} |\boldsymbol{\Sigma}, \mathbf{v}^2, \mathbf{R}_0 \right) = \begin{pmatrix} \mathbf{1}\mu_1 \\ \mathbf{1}\mu_2 \end{pmatrix}, \tag{58}$$

and

$$Var\left(\left[\begin{array}{c}\mathbf{y}_1\\ \mathbf{y}_2\end{array}\right]|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0\right)=\mathbf{C}_{Cond}+\mathbf{R}_0\otimes\mathbf{I}=\mathbf{V}_{Cond}. \tag{59}$$

After assigning a flat prior to each of μ_1 and μ_2, standard results give that posterior distribution of the genotypic values given $\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0$ is normal, with mean vector

$$E\left(\left[\begin{array}{c}\mathbf{g}_1\\ \mathbf{g}_2\end{array}\right]|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0,\mathbf{y}\right)=\mathbf{C}_{Cond}\mathbf{V}_{Cond}^{-1}\left(\left[\begin{array}{c}\mathbf{y}_1-\mathbf{1}\widetilde{\mu}_1\\ \mathbf{y}_2-\mathbf{1}\widetilde{\mu}_2\end{array}\right]\right)=\left[\begin{array}{c}\widetilde{\mathbf{g}}_1\\ \widetilde{\mathbf{g}}_2\end{array}\right], \tag{60}$$

where

$$\begin{aligned}\left[\begin{array}{c}\widetilde{\mu}_1\\ \widetilde{\mu}_2\end{array}\right] &= \left\{\left[\begin{array}{cc}\mathbf{1}' & 0\\ 0 & \mathbf{1}'\end{array}\right]\mathbf{V}_{Cond}^{-1}\left[\begin{array}{cc}\mathbf{1} & 0\\ 0 & \mathbf{1}\end{array}\right]\right\}^{-1}\\ &\quad\times\left\{\left[\begin{array}{cc}\mathbf{1}' & 0\\ 0 & \mathbf{1}'\end{array}\right]\mathbf{V}_{Cond}^{-1}\left[\begin{array}{c}\mathbf{y}_1\\ \mathbf{y}_2\end{array}\right]\right\}^{-1}.\end{aligned}$$

Further (Henderson 1975)

$$\begin{aligned}&Var\left(\left[\begin{array}{c}\mathbf{g}_1\\ \mathbf{g}_2\end{array}\right]|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0,\mathbf{y}\right)\\ &= \mathbf{C}_{Cond}-\mathbf{C}_{Cond}\mathbf{V}_{Cond}^{-1}\mathbf{C}_{Cond}+\mathbf{C}_{Cond}\mathbf{V}_{Cond}^{-1}\mathbf{1}\left(\mathbf{1}'\mathbf{V}_{Cond}^{-1}\mathbf{1}\right)\mathbf{1}'\mathbf{V}_{Cond}^{-1}\mathbf{C}_{Cond}\\ &= \mathbf{C}_{Cond}-\mathbf{C}_{Cond}\left[\mathbf{V}_{Cond}^{-1}-\mathbf{V}_{Cond}^{-1}\mathbf{1}\left(\mathbf{1}'\mathbf{V}_{Cond}^{-1}\mathbf{1}\right)\mathbf{1}'\mathbf{V}_{Cond}^{-1}\right]\mathbf{C}_{Cond}.\end{aligned} \tag{61}$$

Hence, draws from the conditional posterior distribution of $\mathbf{g}=\left[\begin{array}{cc}\mathbf{g}_1 & \mathbf{g}_2\end{array}\right]'$ given $\boldsymbol{\Sigma},\mathbf{v}^2$ and $\mathbf{R}_0$ can be obtained by sampling from a multivariate normal distribution with mean vector (60) and covariance matrix (61).

Assuming that, given $\boldsymbol{\Sigma},\mathbf{v}^2$ and $\mathbf{R}_0$, the vector $\left[\begin{array}{cccc}\boldsymbol{\beta}_1^{*\prime} & \boldsymbol{\beta}_2^{*\prime} & \mathbf{g}_1' & \mathbf{g}_2'\end{array}\right]'$ has a multivariate normal distribution, and let $\boldsymbol{\beta}^{*\prime}=\left[\begin{array}{cc}\boldsymbol{\beta}_1^{*\prime} & \boldsymbol{\beta}_2^{*\prime}\end{array}\right]'$. Hence

$$\begin{aligned}E\left(\boldsymbol{\beta}^*|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0,\mathbf{y}\right) &= E_{\mathbf{g}|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0,\mathbf{y}}\left[\boldsymbol{\beta}^*|\mathbf{g},\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0\right]\\ &= E_{\mathbf{g}|\boldsymbol{\Sigma},\mathbf{v}^2,\mathbf{R}_0,\mathbf{y}}\left[Cov\left(\boldsymbol{\beta}^*,\mathbf{g}'\right)\left(\boldsymbol{\Sigma}\otimes\mathbf{XDX}'\right)^{-1}\mathbf{g}\right]\\ &= Cov\left(\boldsymbol{\beta}^*,\mathbf{g}'\right)\left(\boldsymbol{\Sigma}\otimes\mathbf{XDX}'\right)^{-1}\widetilde{\mathbf{g}}=\widetilde{\boldsymbol{\beta}}^*.\end{aligned} \tag{62}$$

Now,

$$\begin{aligned}Cov\left(\boldsymbol{\beta}^*,\mathbf{g}'\right) &= Cov\left(\left[\begin{array}{c}\boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\end{array}\right],\left[\begin{array}{cc}\mathbf{g}_1' & \mathbf{g}_2'\end{array}\right]\right)\\ &= \mathbf{B}\otimes\mathbf{X}',\end{aligned} \tag{63}$$

so

$$\widetilde{\boldsymbol{\beta}}^{*} = \left(\mathbf{B} \otimes \mathbf{X}'\right)\left(\boldsymbol{\Sigma} \otimes \mathbf{X}\mathbf{D}\mathbf{X}'\right)^{-1} = \left(\mathbf{B}\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}\mathbf{D}\mathbf{X}'\right)\widetilde{\mathbf{g}} \tag{64}$$

Similarly

$$\begin{aligned} Var\left(\boldsymbol{\beta}^{*}|\boldsymbol{\Sigma}, \mathbf{v}^{2}, \mathbf{R}_{0}, \mathbf{y}\right) &= Var_{\mathbf{g}|\boldsymbol{\Sigma},\mathbf{v}^{2},\mathbf{R}_{0},\mathbf{y}}\left[E\left(\boldsymbol{\beta}^{*}|\mathbf{g}, \boldsymbol{\Sigma}, \mathbf{v}^{2}, \mathbf{R}_{0}\right)\right] \\ &\quad + E_{\mathbf{g}|\boldsymbol{\Sigma},\mathbf{v}^{2},\mathbf{R}_{0},\mathbf{y}} Var\left[\boldsymbol{\beta}^{*}|\mathbf{g}, \boldsymbol{\Sigma}, \mathbf{v}^{2}, \mathbf{R}_{0}\right] \\ &= Var_{\mathbf{g}|\boldsymbol{\Sigma},\mathbf{v}^{2},\mathbf{R}_{0},\mathbf{y}}\left[\left(\mathbf{B}\boldsymbol{\Sigma}^{-1} \otimes \mathbf{X}'\mathbf{X}\mathbf{D}\mathbf{X}'\right)\mathbf{g}\right] \\ &\quad + \left(\mathbf{B} \otimes \mathbf{I}\right)\left(\boldsymbol{\Sigma} \otimes \mathbf{X}\mathbf{D}\mathbf{X}'\right)^{-1}\left(\mathbf{B} \otimes \mathbf{I}\right) \\ &= \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B} \otimes \left(\mathbf{X}\mathbf{D}\mathbf{X}'\right)^{-1}. \end{aligned} \tag{65}$$

13.5 Appendix E: A conditional posterior mode approximation to marker effects

In spite of important advances in high-throughput computing, routine genetic evaluation of plants and animals is seldom done with MCMC methods. As an alternative to MCMC, we describe an iterative algorithm that produces point estimates of marker effects (and of linear functions thereof) and approximate measures of uncertainty in a computationally simpler manner. The algorithm uses a re-weighted set of linear "mixed model equations", for which extremely efficient solvers exist. It is assumed that "good" estimates of $\mathbf{R}_0$ (the residual covariance matrix) and of $\mathbf{B}$ (the $T \times T$ variance-covariance matrix of markers effects) are available. From (8) $\boldsymbol{\Sigma} = \mathbf{B}/\left[4\left(T+1\right)\right]$, e.g., for $T = 3$ then $\boldsymbol{\Sigma} = \mathbf{B}/16$; hence, an assessment of the scale matrix of the MLAP distribution is easily available.

We make use of (2) and of (7) but employ the "markers within trait" representation given in (4). Letting $\boldsymbol{\theta} = \left(\boldsymbol{\mu}', \boldsymbol{\beta}'\right)'$, the logarithm of the joint (conditionally on the dispersion matrices) posterior density of location effects, apart from a constant, is

$$\begin{aligned} &\log\left[p(\boldsymbol{\theta}|\mathbf{R}_0, \boldsymbol{\Sigma}, DATA)\right] \\ &= -\frac{1}{2}\left\{\mathbf{y}^{*} - \left(\mathbf{I}_3 \otimes \mathbf{1}_N\right)\boldsymbol{\mu} - \left(\mathbf{I}_3 \otimes \mathbf{X}\right)\boldsymbol{\beta}^{*}\right\}'\left(\mathbf{R}_0^{-1} \otimes \mathbf{I}_N\right)\left\{\mathbf{y}^{*} - \left(\mathbf{I}_3 \otimes \mathbf{1}_N\right)\boldsymbol{\mu} - \left(\mathbf{I}_3 \otimes \mathbf{X}\right)\boldsymbol{\beta}^{*}\right\} \\ &\quad -\frac{1}{2}\sum_{j=1}^{p}\sqrt{\boldsymbol{\beta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j} = L(\boldsymbol{\theta}) \end{aligned} \tag{66}$$

Let

$$L(\boldsymbol{\theta}) = L_{lik}(\boldsymbol{\mu}, \boldsymbol{\beta}^{*}) + L_{prior}(\boldsymbol{\beta}), \tag{67}$$

where $L_{lik}(\boldsymbol{\mu}, \boldsymbol{\beta}^*)$ and $L_{prior}(\boldsymbol{\beta})$ are the two terms in (66). Then

$$\frac{\partial L_{lik}(\boldsymbol{\mu}, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\mu}} = \left(\mathbf{R}_0^{-1} \otimes \mathbf{1}_N'\right) \left[\mathbf{y}^* - \left(\mathbf{I}_3 \otimes \mathbf{1}_N\right) \boldsymbol{\mu} - \left(\mathbf{I}_3 \otimes \mathbf{X}\right) \boldsymbol{\beta}^*\right]. \tag{68}$$

and

$$\frac{\partial L_{lik}(\boldsymbol{\mu}, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \left(\mathbf{R}_0^{-1} \otimes \mathbf{X}'\right) \left[\mathbf{y}^* - \left(\mathbf{I}_3 \otimes \mathbf{1}_N\right) \boldsymbol{\mu} - \left(\mathbf{I}_3 \otimes \mathbf{X}\right) \boldsymbol{\beta}^*\right] \tag{69}$$

Observe now that the relationship between $\boldsymbol{\beta}$ and marker effects sorted within traits ($\boldsymbol{\beta}^*$) can be expressed as $\boldsymbol{\beta} = \mathbf{L}\boldsymbol{\beta}^*$ where $\mathbf{L}$ is a $3p \times 3p$ non-singular matrix of elementary operators that rearrange rows and columns. For example, for $T = 3$ and $p = 2$ and with β_{jt} representing the effect of marker j on trait t,

$$\begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{11}^* \\ \beta_{21}^* \\ \beta_{12}^* \\ \beta_{22}^* \\ \beta_{13}^* \\ \beta_{23}^* \end{bmatrix}. \tag{70}$$

Since $\mathbf{L}$ is a matrix of elementary operators, $\mathbf{L}^{-1} = \mathbf{L}'$ (orthogonality) and $\boldsymbol{\beta}^* = \mathbf{L}'\boldsymbol{\beta}$; the absolute value of the Jacobian of the transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\beta}^*$ is equal to 1. The contribution of the prior to the gradient for marker effects is then

$$\frac{\partial}{\partial \boldsymbol{\beta}} L_{prior}(\boldsymbol{\beta}) = -\boldsymbol{\Sigma}^{-1} \frac{1}{m_j} \boldsymbol{\beta}_j = \begin{bmatrix} \frac{1}{m_j}\left(\Sigma^{11}\beta_{j1} + \Sigma^{12}\beta_{j2} + \Sigma^{13}\beta_{j3}\right) \\ \frac{1}{m_j}\left(\Sigma^{21}\beta_{j1} + \Sigma^{22}\beta_{j2} + \Sigma^{23}\beta_{j3}\right) \\ \frac{1}{m_j}\left(\Sigma^{31}\beta_{j1} + \Sigma^{32}\beta_{j2} + \Sigma^{33}\beta_{j3}\right) \end{bmatrix}; \; j = 1, 2..., p. \tag{71}$$

where $m_j = 2\sqrt{\boldsymbol{\beta}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_j}$ is proportional to the the Mahalanobis distance of $\boldsymbol{\beta}_j$ away from $(0, 0, 0)$ for $T = 3$. Hence, the $3p \times 1$ vector of derivatives with respect to all marker effects, sorted by traits within individuals is

$$\frac{\partial}{\partial \boldsymbol{\beta}} L_{prior}(\boldsymbol{\beta}) = -\begin{bmatrix} \boldsymbol{\Sigma}^{-1} m_1 \boldsymbol{\beta}_1 \\ \boldsymbol{\Sigma}^{-1} m_2 \boldsymbol{\beta}_2 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\Sigma}^{-1} m_p \boldsymbol{\beta}_p \end{bmatrix} = -\left(\mathbf{M}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{\beta}, \tag{72}$$

where $\mathbf{M} = Diag\{m_j\}$ is a $p \times p$ diagonal matrix with typical element m_j. Rearranging the differentials such that the sorting is by markers within traits

$$\frac{\partial}{\partial \boldsymbol{\beta}^*} L_{prior}(\boldsymbol{\beta}^*) = -\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{M}^{-1}\right) \boldsymbol{\beta}^*. \tag{73}$$

Collecting (69) and (73),

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}^*} L(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\beta}^*} L_{lik}(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\beta}^*} L_{prior}(\boldsymbol{\beta}^*) \\ &= \left(\mathbf{R}_0^{-1} \otimes \mathbf{X}'\right)\left[\mathbf{y}^* - \left(\mathbf{I}_3 \otimes \mathbf{1}_N\right)\boldsymbol{\mu} - \left(\mathbf{I}_3 \otimes \mathbf{X}\right)\boldsymbol{\beta}^*\right] - \left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{M}^{-1}\right)\boldsymbol{\beta}^*, \end{aligned} \tag{74}$$

Setting (68) and (74) simultaneously to $\mathbf{0}$ produces the system of equations (not explicit)

$$\begin{bmatrix} \mathbf{R}_0^{-1} \otimes N & \mathbf{R}_0^{-1} \otimes \mathbf{1}'_N\mathbf{X} \\ \mathbf{R}_0^{-1} \otimes \mathbf{X}'\mathbf{1}_N & \left(\mathbf{I}_3 \otimes \mathbf{X}'\mathbf{X}\right) + \left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{M}^{-1}\right) \end{bmatrix} \begin{bmatrix} \overline{\boldsymbol{\mu}} \\ \overline{\boldsymbol{\beta}}^* \end{bmatrix} = \begin{bmatrix} \left(\mathbf{R}_0^{-1} \otimes \mathbf{1}'_N\right)\mathbf{y}^* \\ \left(\mathbf{R}_0^{-1} \otimes \mathbf{X}'\right)\mathbf{y}^* \end{bmatrix}. \tag{75}$$

Expanding the equations above for $T = 3$ yields

$$\begin{bmatrix} r^{11}N & r^{12}N & r^{13}N & r^{11}\mathbf{1}'_N\mathbf{X} & r^{12}\mathbf{1}'_N\mathbf{X} & r^{13}\mathbf{1}'_N\mathbf{X} \\ r^{21}N & r^{22}N & r^{23}N & r^{21}\mathbf{1}'_N\mathbf{X} & r^{22}\mathbf{1}'_N\mathbf{X} & r^{23}\mathbf{1}'_N\mathbf{X} \\ r^{31}N & r^{32}N & r^{33}N & r^{31}\mathbf{1}'_N\mathbf{X} & r^{32}\mathbf{1}'_N\mathbf{X} & r^{33}\mathbf{1}'_N\mathbf{X} \\ r^{11}\mathbf{X}'\mathbf{1}_N & r^{12}\mathbf{X}'\mathbf{1}_N & r^{13}\mathbf{X}'\mathbf{1}_N & \mathbf{X}'r^{11}\mathbf{X}+\Sigma^{11}\mathbf{M}^{-1} & \mathbf{X}'r^{12}\mathbf{X}+\Sigma^{12}\mathbf{M}^{-1} & \mathbf{X}'r^{13}\mathbf{X}+\Sigma^{13}\mathbf{M}^{-1} \\ r^{21}\mathbf{X}'\mathbf{1}_N & r^{22}\mathbf{X}'\mathbf{1}_N & r^{23}\mathbf{X}'\mathbf{1}_N & \mathbf{X}'r^{12}\mathbf{X}+\Sigma^{21}\mathbf{M}^{-1} & \mathbf{X}'r^{22}\mathbf{X}+\Sigma^{22}\mathbf{M}^{-1} & \mathbf{X}'r^{23}\mathbf{X}+\Sigma^{23}\mathbf{M}^{-1} \\ r^{31}\mathbf{X}'\mathbf{1}_N & r^{32}\mathbf{X}'\mathbf{1}_N & r^{33}\mathbf{X}'\mathbf{1}_N & \mathbf{X}'r^{31}\mathbf{X}+\Sigma^{31}\mathbf{M}^{-1} & \mathbf{X}'r^{32}\mathbf{X}+\Sigma^{32}\mathbf{M}^{-1} & \mathbf{X}'r^{33}\mathbf{X}+\Sigma^{33}\mathbf{M}^{-1} \end{bmatrix}^{[b]} \times$$

$$\begin{bmatrix} \overline{\mu}_1 \\ \overline{\mu}_2 \\ \overline{\mu}_3 \\ \overline{\boldsymbol{\beta}}_1^* \\ \overline{\boldsymbol{\beta}}_2^* \\ \overline{\boldsymbol{\beta}}_3^* \end{bmatrix}^{[b+1]} = \begin{bmatrix} \mathbf{1}'_N(r^{11}\mathbf{y}_1^* + r^{12}\mathbf{y}_2^* + r^{13}\mathbf{y}_3^*) \\ \mathbf{1}'_N\left(r^{21}\mathbf{y}_1^* + r^{22}\mathbf{y}_2^* + r^{23}\mathbf{y}_3^*\right) \\ \mathbf{1}'_N\left(r^{31}\mathbf{y}_1^* + r^{32}\mathbf{y}_2^* + r^{33}\mathbf{y}_3^*\right) \\ \mathbf{X}'\left(r^{11}\mathbf{y}_1^* + r^{12}\mathbf{y}_2^* + r^{13}\mathbf{y}_3^*\right) \\ \mathbf{X}'\left(r^{21}\mathbf{y}_1^* + r^{22}\mathbf{y}_2^* + r^{23}\mathbf{y}_3^*\right) \\ \mathbf{X}'\left(r^{31}\mathbf{y}_1^* + r^{32}\mathbf{y}_2^* + r^{33}\mathbf{y}_3^*\right) \end{bmatrix}^{[b]}, \tag{76}$$

where b is iterate round. Matrix $\mathbf{M} = Diag(m_j)$ changes at every round of iteration, so the system needs to be reconstituted repeatedly. Marker effects producing small values of the Mahalanobis distance away from 0 result in tiny $m-$values and, consequently, $\mathbf{M}^{-1}$ will have large diagonal elements. Hence, vectors of markers with weak effects are more strongly shrunk towards the 0 coordinate than those having strong effects in at least one trait

The variance-covariance matrix of the conditional posterior distribution can be approximated

as

$$Var\left(\begin{bmatrix}\mu_1\\ \mu_1\\ \mu_1\\ \boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\\ \boldsymbol{\beta}_3^*\end{bmatrix}|\mathbf{R}_0,\boldsymbol{\Sigma},DATA\right)\approx\mathbf{K}_{[\infty]}^{-1}, \tag{77}$$

with ∞ indicating parameters evaluated at converged values, assuming that convergence has been attained at a hopefully global mode.

13.6 Appendix F: Treatment of missing data

Let a multi-trait data point $(T\times 1)$ on individual i be represented as

$$\mathbf{y}_i^{complete}=\left(\mathbf{y}_i^{miss},\mathbf{y}_i^{obs}\right), \tag{78}$$

where "$miss$" denotes a missing record, e.g., if $T=2$, a record could be missing for trait 1 or for trait 2; $\mathbf{y}_i^{obs}$ represents the phenotypes for the traits observed in individual i. The posterior predictive distribution of $\mathbf{y}_i^{miss}$ has density

$$p\left(\mathbf{y}_i^{miss}|\mathbf{y},\mathbf{R}_0,\Sigma\right)=\int_{\Re_{\boldsymbol{\mu}}}\int_{\Re_{\boldsymbol{\beta}}}p\left(\mathbf{y}_i^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_i^{obs}\right)p\left(\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R_0},\boldsymbol{\Sigma}|\mathbf{y}\right)d\boldsymbol{\mu}d\boldsymbol{\beta}d\mathbf{R_0}d\boldsymbol{\Sigma}, \tag{79}$$

provided that data points in i are conditionally (given $\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0$) independent of any other i' in the sample, and with $\mathbf{y}$ being all observed data. The preceding formulae implies that $\mathbf{y}_i^{miss}$ can be imputed by sampling $\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R_0},\boldsymbol{\Sigma}$ from their posterior distribution and then drawing from

$$\mathbf{y}_i^{miss}|\mathbf{y},\mathbf{R}_0,\Sigma\sim N\left(E\left(\mathbf{y}_i^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_i^{obs}\right),Var\left(\mathbf{y}_i^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_i^{obs}\right)\right) \tag{80}$$

Since the sampling model is normal, for $T=3$ one has

$$\begin{aligned}E\left(\mathbf{y}_i^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_i^{obs}\right) &= \begin{bmatrix}\delta_1\mu_1\\ \delta_2\mu_2\\ \delta_2\mu_3\end{bmatrix}+\begin{bmatrix}\delta_1\mathbf{x}_i' & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \delta_2\mathbf{x}_i' & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \delta_3\mathbf{x}_i'\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1^*\\ \boldsymbol{\beta}_2^*\\ \boldsymbol{\beta}_3^*\end{bmatrix}\\ &+\mathbf{R}_0^{[miss,obs]}\left(\mathbf{R}_0^{[obs,obs]}\right)^{-1}\mathbf{e}^{[obs]},\end{aligned} \tag{81}$$

where $\delta_1,\delta_2,\delta_3$ take the value 1 when a given trait is missing in case i, or denote "exclude from formula" otherwise; $\mathbf{R}_0^{[obs,obs]}$ is the part of $\mathbf{R}_0$ pertaining to observed phenotypes for case i, and

$\mathbf{R}_0^{[miss,obs]}$ is the submatrix of residual covariances between missing and observed traits. Further,

$$Var\left(\mathbf{y}_i^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_i^{obs}\right)=\mathbf{R}_0^{[miss,miss]}-\mathbf{R}_0^{[miss,obs]}\left(\mathbf{R}_0^{[obs,obs]}\right)^{-1}\mathbf{R}_0^{[obs,miss]} \tag{82}$$

For example, let $T=3$ and suppose that trait 1 is missing in case 250 but that traits 2 and 3 have been recorded; here

$$E\left(y_{250}^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_{250}^{obs}\right)=\mu_1+\mathbf{x}'_{250}\boldsymbol{\beta}_1^*+\begin{bmatrix} r_{12} & r_{13}\end{bmatrix}\begin{bmatrix} r_{22} & r_{23}\\ r_{32} & r_{33}\end{bmatrix}^{-1}\begin{bmatrix} e_{2,250}\\ e_{3,250}\end{bmatrix}, \tag{83}$$

and

$$Var\left(y_{250}^{miss}|\boldsymbol{\mu},\boldsymbol{\beta},\mathbf{R}_0,\mathbf{y}_{250}^{obs}\right)=r_{11}-\begin{bmatrix} r_{12} & r_{13}\end{bmatrix}\begin{bmatrix} r_{22} & r_{23}\\ r_{32} & r_{33}\end{bmatrix}^{-1}\begin{bmatrix} r_{12}\\ r_{13}\end{bmatrix}. \tag{84}$$

In the MCMC algorithm, missing data are sampled independently across cases, but dependently within case by addressing the pattern of missingness peculiar to each observation. Samples for missing observations can be used to estimate predictive distributions for the missing data (Gelfand et al. 1992; Sorensen and Gianola 2002; Gelman et al. 2014).

14 Legend for Supplemental Figures

Figure S1. Five bivariate Laplace densities.

Figure S2. Double exponential versus marginal (from bivariate Laplace) versus normal densities

Figure S3. Normalized densities of mixing variable in MCMC algorithm

Figure S4. Shrinkage factor: mean trait 1

Figure S5. Shrinkage factor: mean trait 2

Figure S6. Shrinkage factor: marker 10, trait 1

Figure S7. Shrinkage factor: marker 200, trait 2

Figure S8. Shrinkage factor: R0[1,1]

Figure S9. Shrinkage factor: R0[1,2]

Figure S10. Shrinkage factor: R0[2,2]

Figure S11. Shrinkage factor: SIGMA[1,1]

Figure S12. Shrinkage factor: SIGMA[1,2]

Figure S13. Shrinkage factor: SIGMA[2,2]

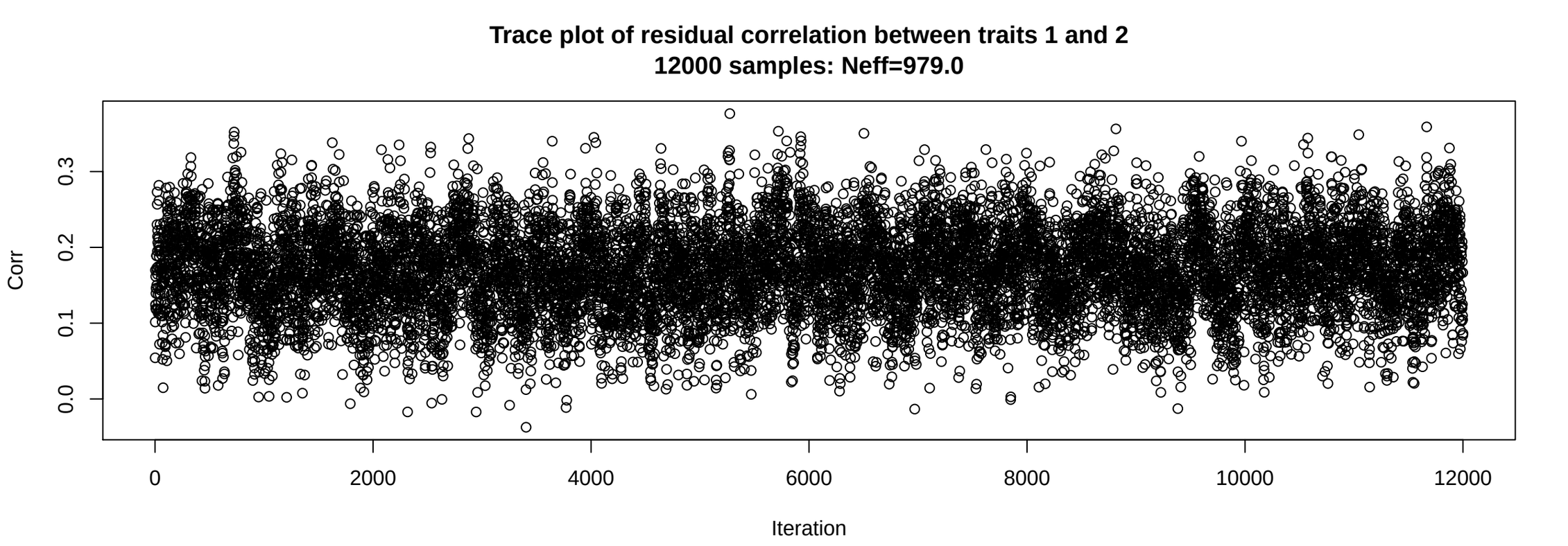
Figure S14. Trace plots of R0[1,1], R0[1,2], R0[2,2]

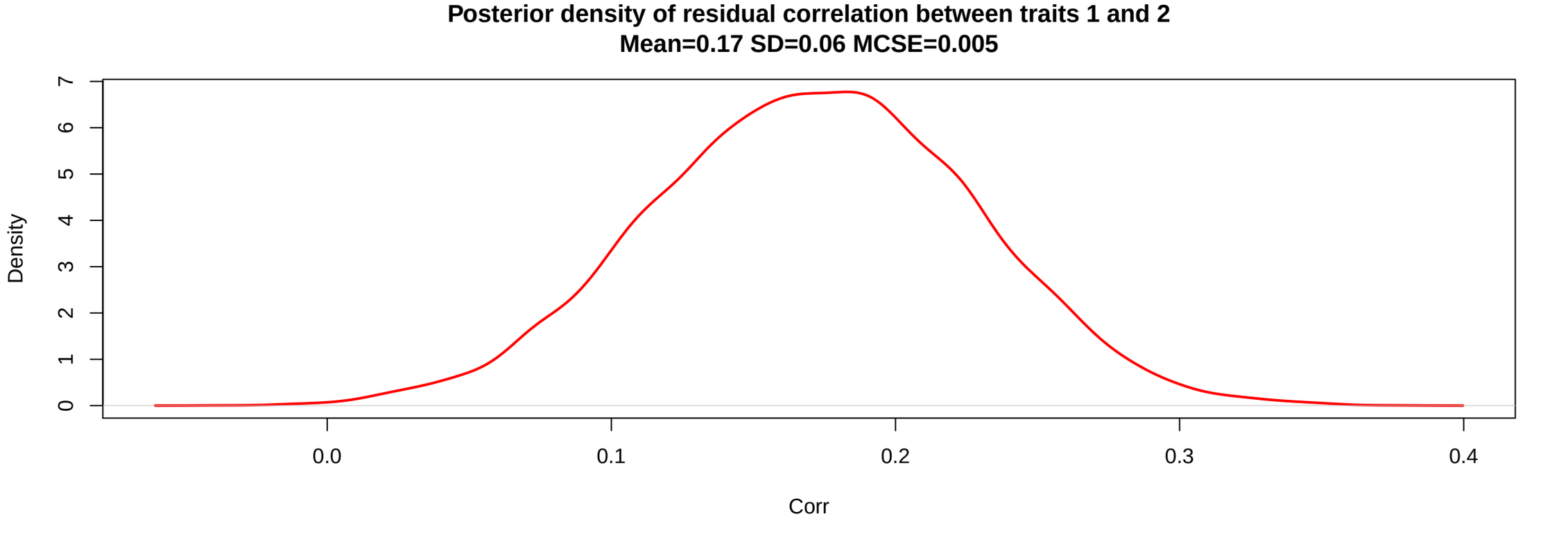
Figure S15. Trace plots of SIGMA[1,1], SIGMA[1,2], SIGMA[2,2]

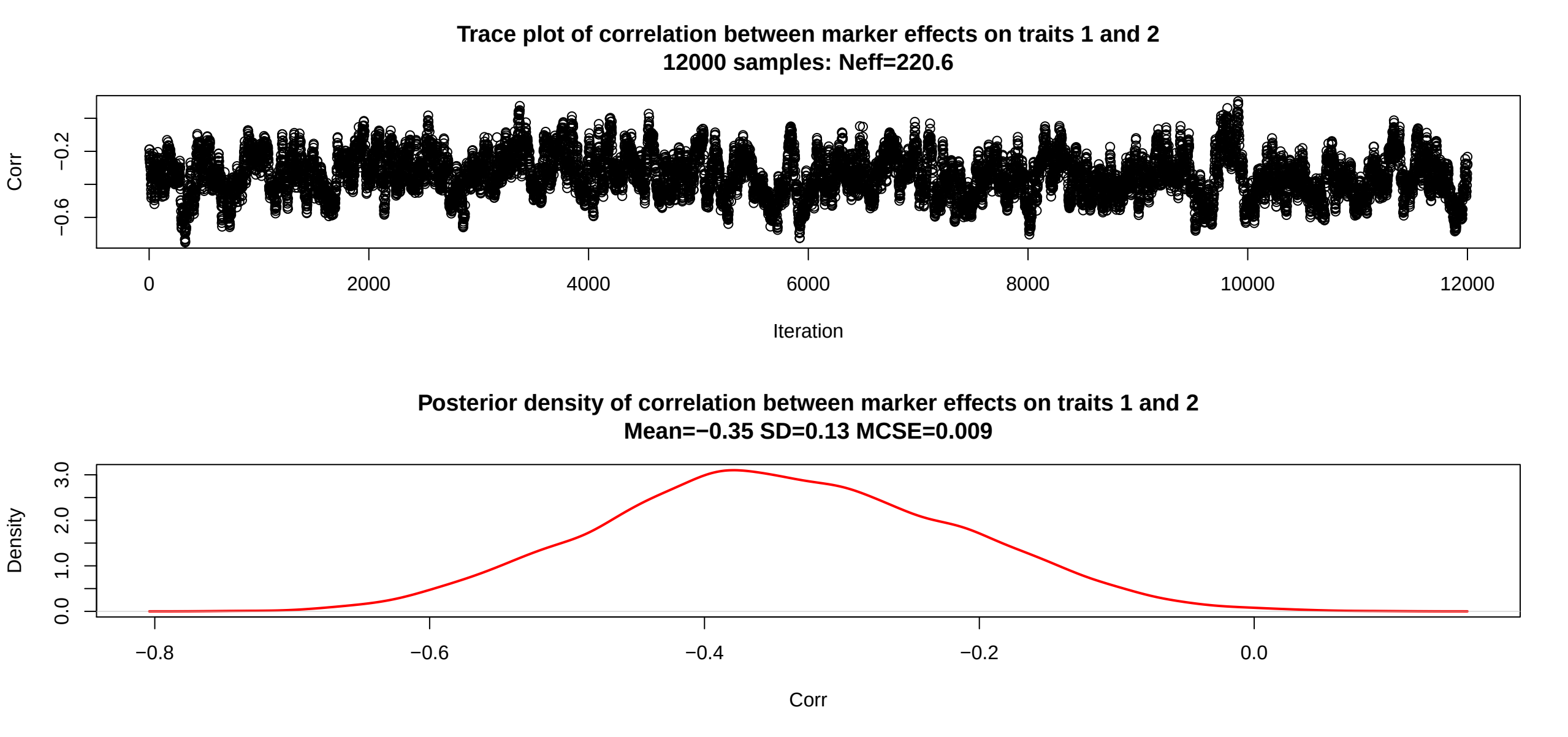
Figure S16. Path to convergence in MAP-MBL (maximum a posteriori-multiple trait Bayesian LASSO)

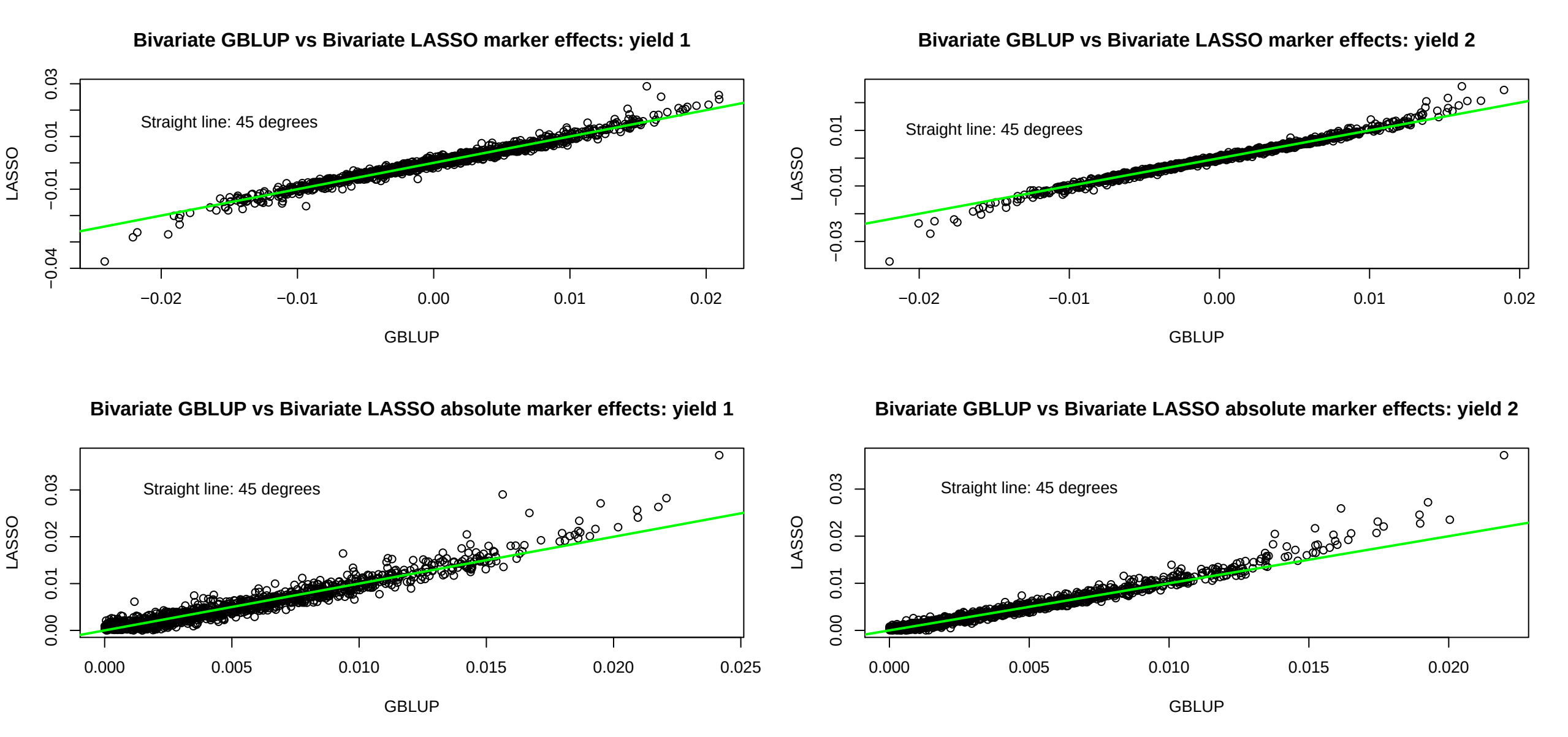
1033 Figure S17. BLUP of marker effects versus MBL posterior means and MAP-MBL solutions

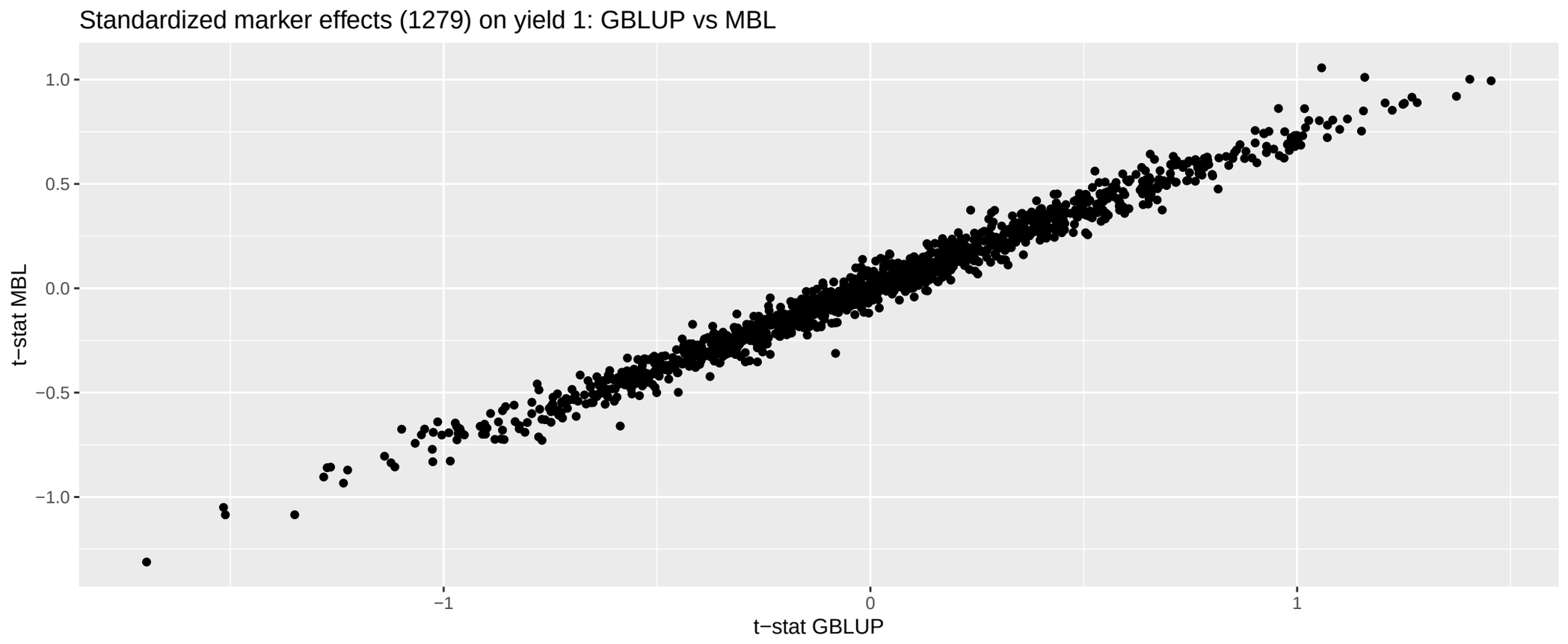
1034 Figure S18. Fitted genetic values: BLUP, MBL and MAP-MBL

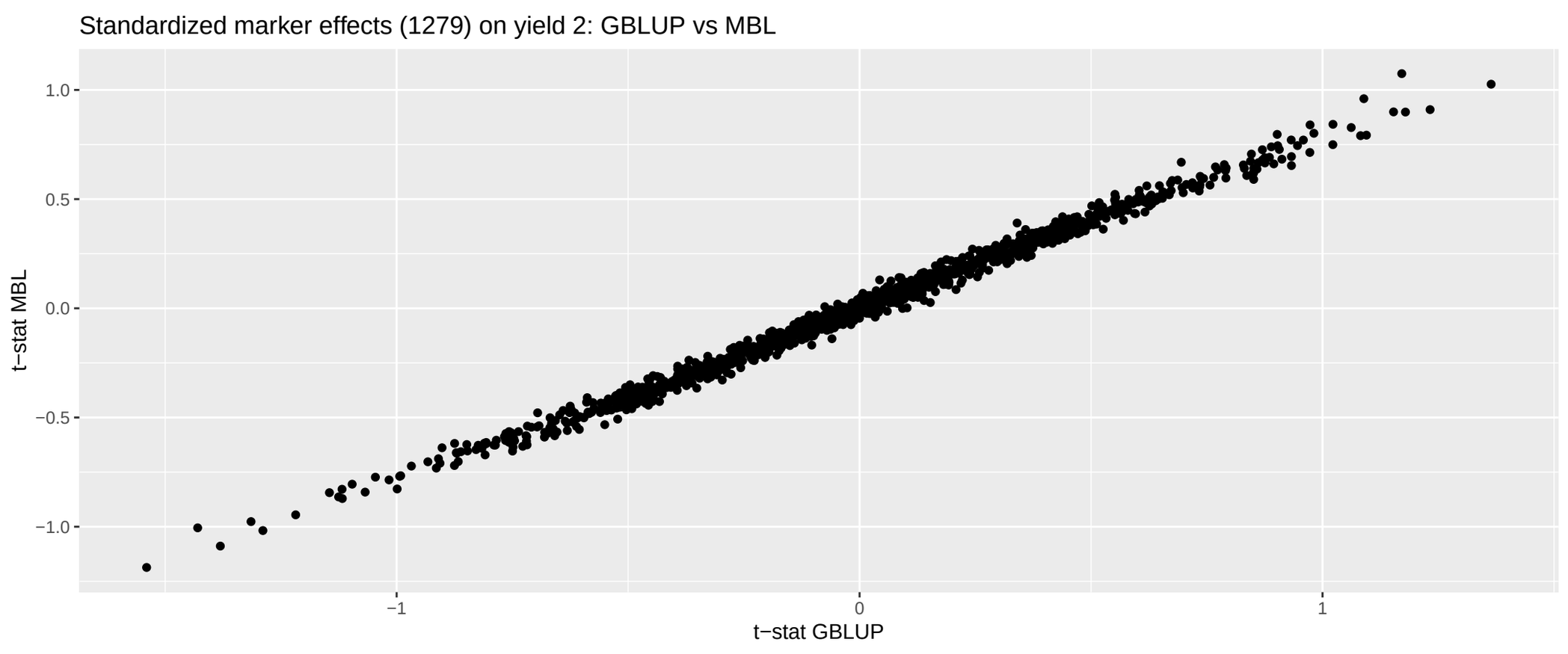

Trace plot of residual correlation between traits 1 and 2
12000 samples: Neff=979.0
Corr
0.0
0.1
0.2
0.3
0
2000
4000
6000
8000
10000
12000
Iteration

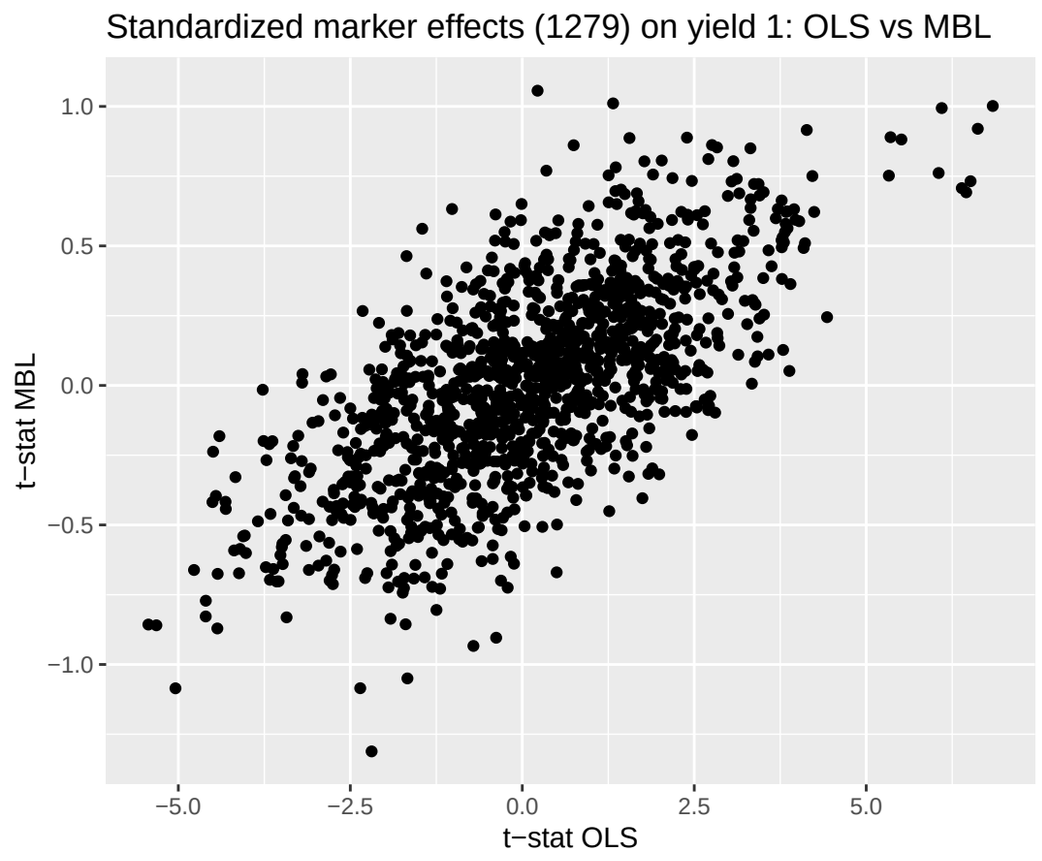

Posterior density of residual correlation between traits 1 and 2
Mean=0.17 SD=0.06 MCSE=0.005
Density
0
1
2
3
4
5
6
7
0.0
0.1
0.2
0.3
0.4
Corr

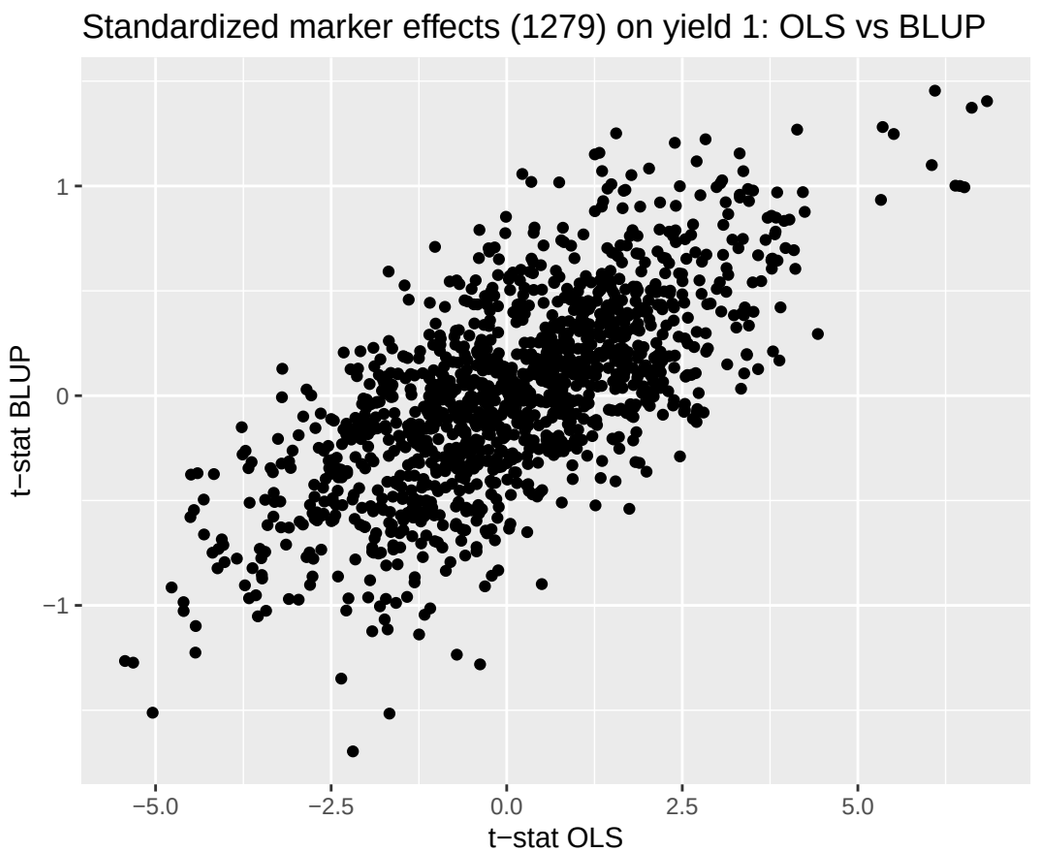

Trace plot of correlation between marker effects on traits 1 and 2
12000 samples: Neff=220.6
Corr
−0.6
−0.2
0
2000
4000
6000
8000
10000
12000
Iteration
Posterior density of correlation between marker effects on traits 1 and 2
Mean=−0.35 SD=0.13 MCSE=0.009
Density
0.0
1.0
2.0
3.0
−0.8
−0.6
−0.4
−0.2
0.0
Corr

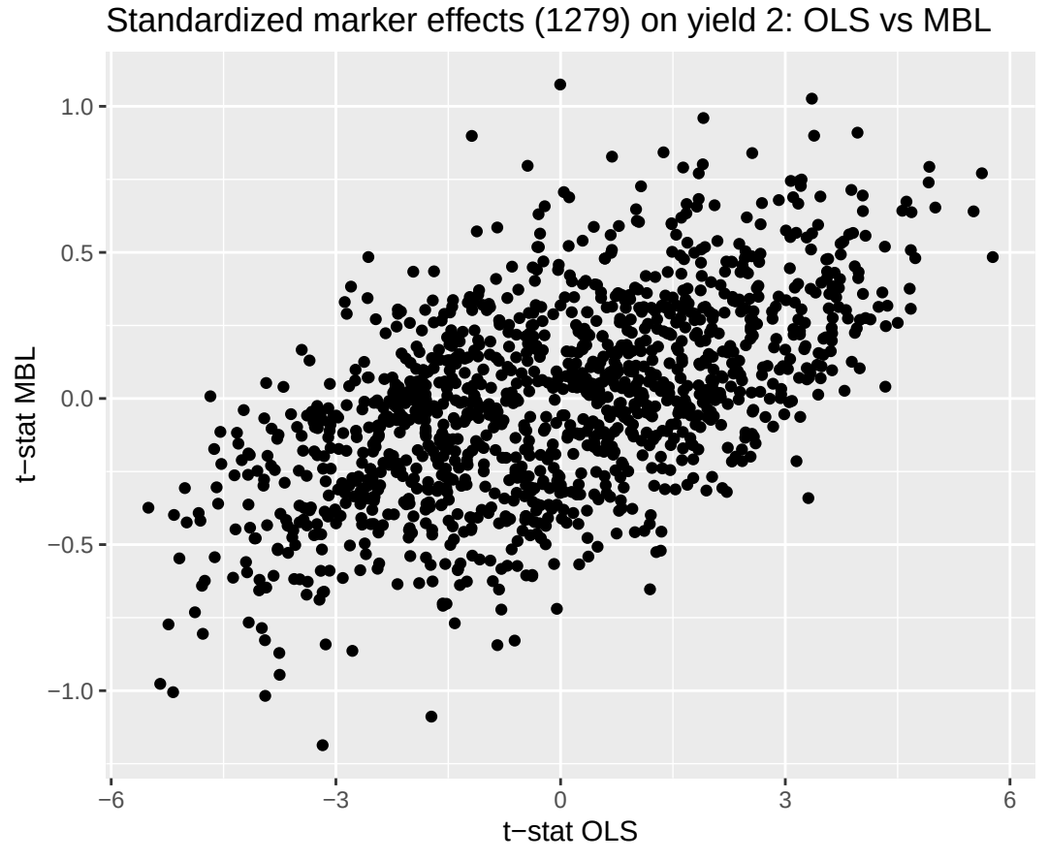

Bivariate GBLUP vs Bivariate LASSO marker effects: yield 1
Straight line: 45 degrees
LASSO
GBLUP
−0.04
−0.01
0.01
0.03
−0.02
−0.01
0.00
0.01
0.02
Bivariate GBLUP vs Bivariate LASSO marker effects: yield 2
Straight line: 45 degrees
LASSO
GBLUP
−0.03
−0.01
0.01
−0.02
−0.01
0.00
0.01
0.02
Bivariate GBLUP vs Bivariate LASSO absolute marker effects: yield 1
Straight line: 45 degrees
LASSO
GBLUP
0.00
0.01
0.02
0.03
0.000
0.005
0.010
0.015
0.020
0.025
Bivariate GBLUP vs Bivariate LASSO absolute marker effects: yield 2
Straight line: 45 degrees
LASSO
GBLUP
0.00
0.01
0.02
0.03
0.000
0.005
0.010
0.015
0.020

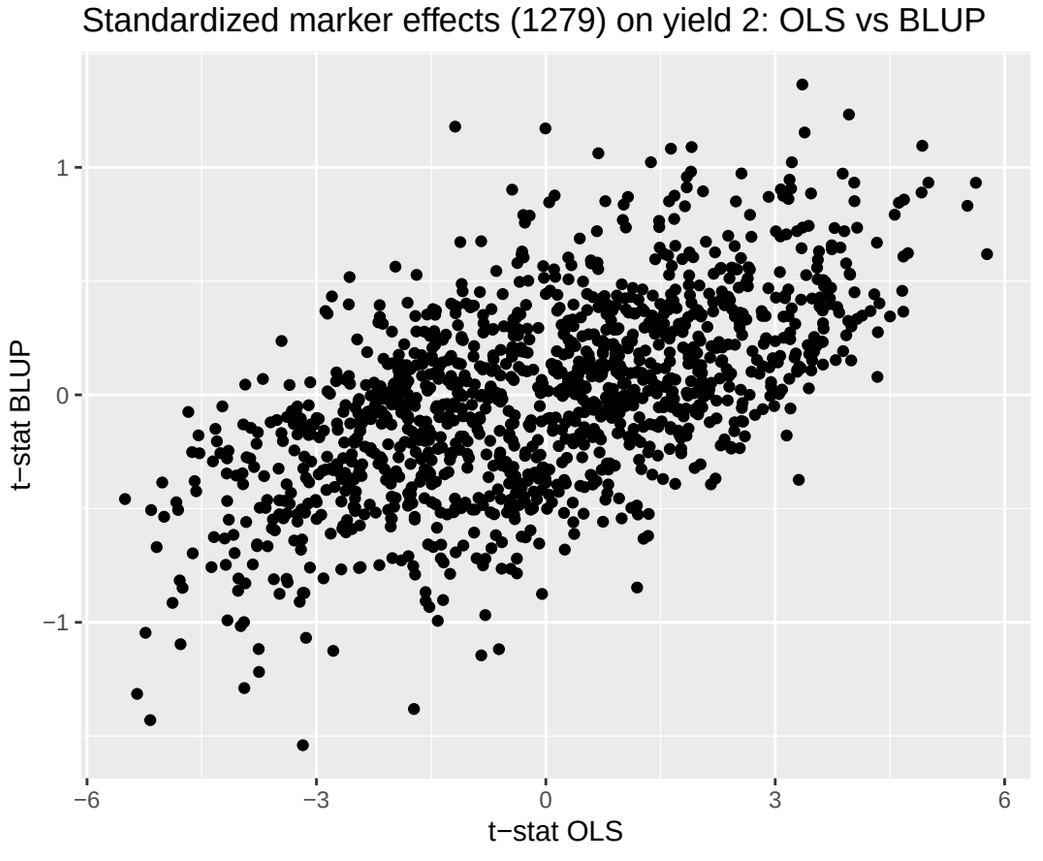

Standardized marker effects (1279) on yield 1: GBLUP vs MBL
t−stat MBL
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
t−stat GBLUP

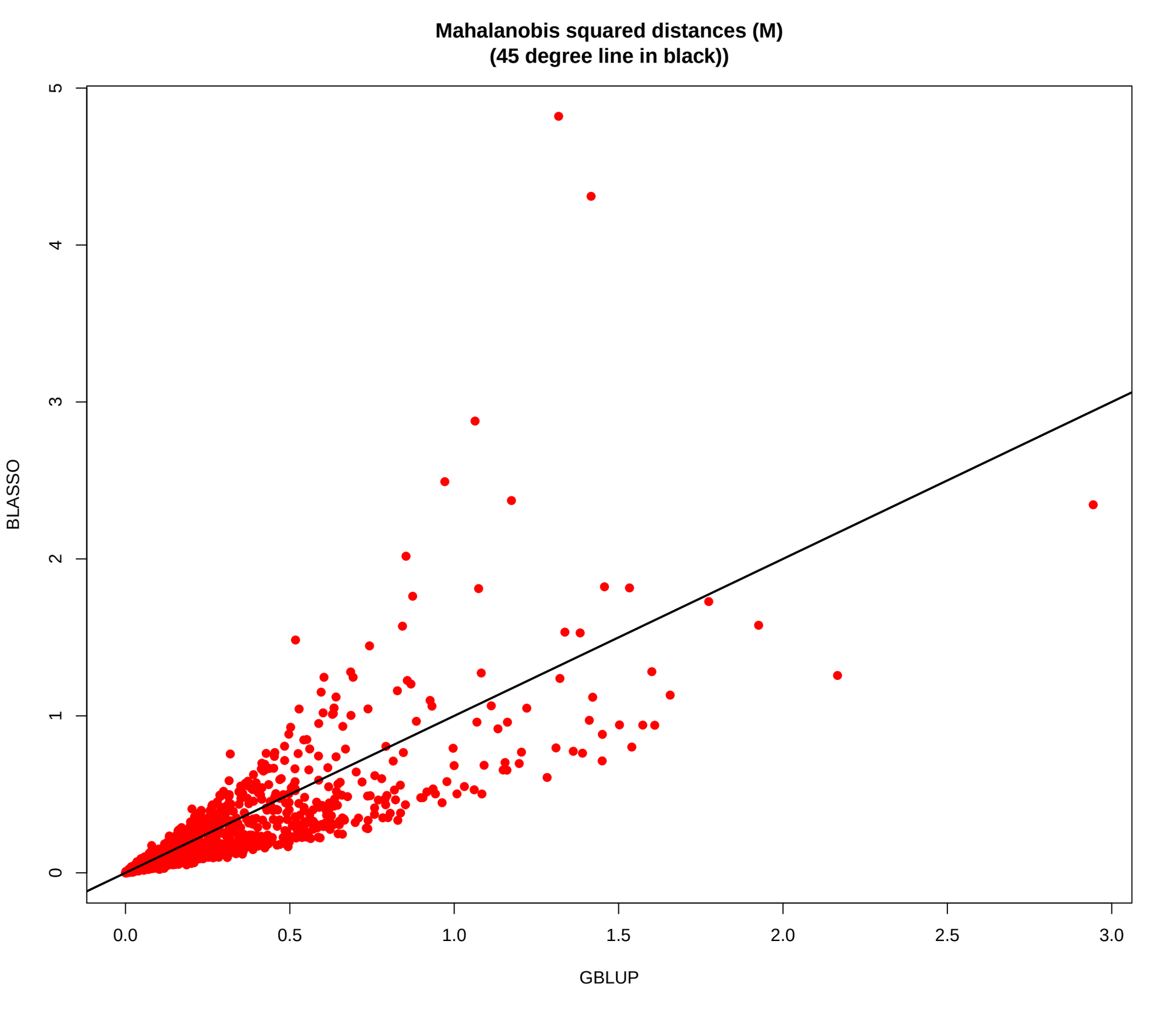

Standardized marker effects (1279) on yield 2: GBLUP vs MBL
t−stat MBL
1.0
0.5
0.0
−0.5
−1.0
−1
0
1
t−stat GBLUP

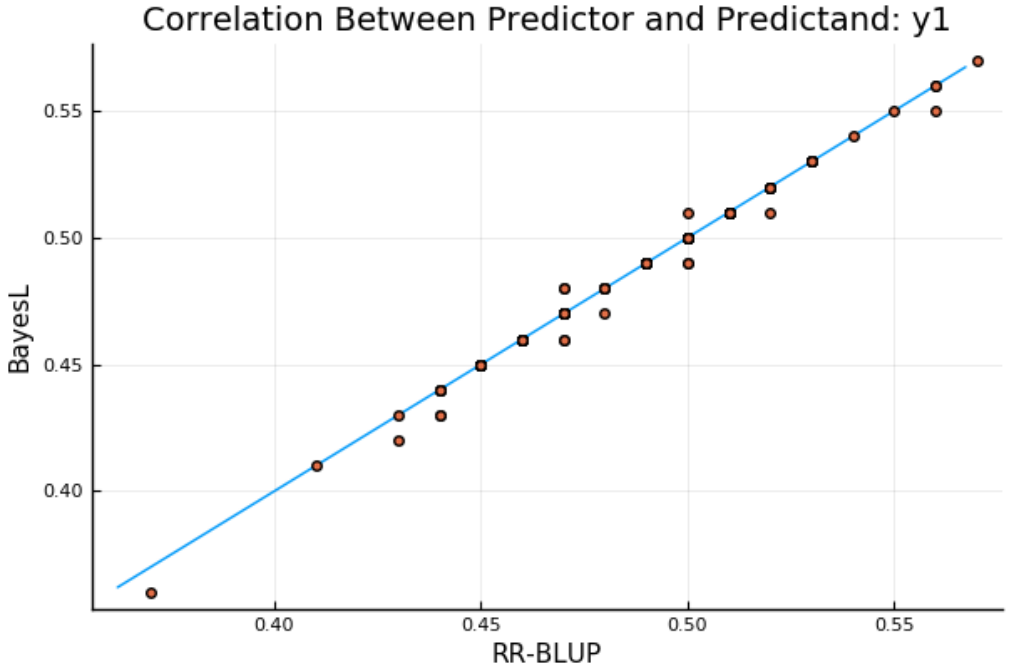

Standardized marker effects (1279) on yield 1: OLS vs MBL
t−stat MBL
t−stat OLS
1.0
0.5
0.0
−0.5
−1.0
−5.0
−2.5
0.0
2.5
5.0

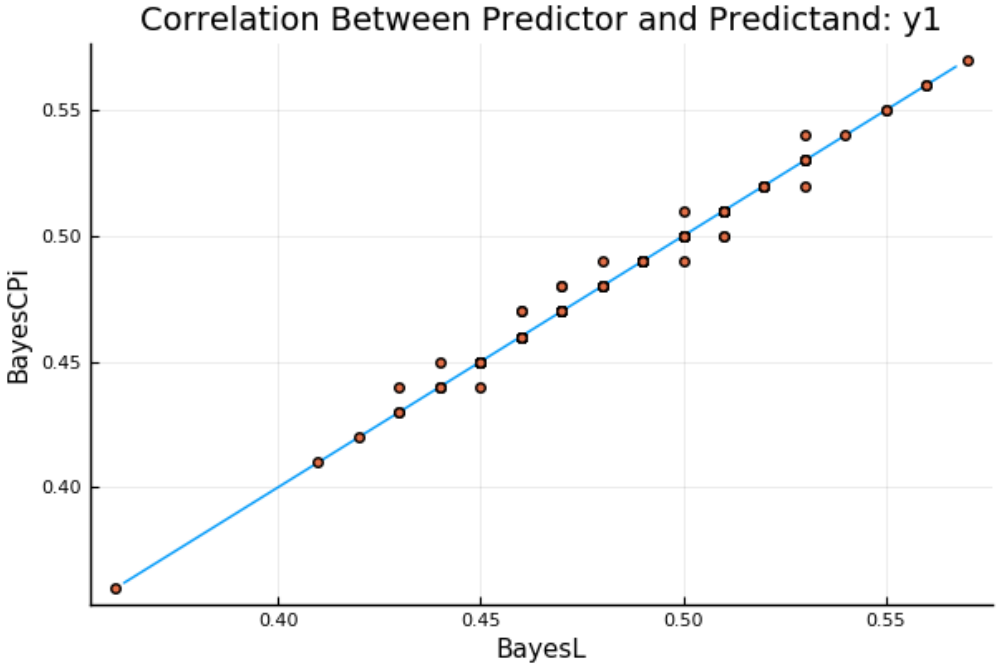

Standardized marker effects (1279) on yield 1: OLS vs BLUP
t−stat BLUP
t−stat OLS
1
0
−1
−5.0
−2.5
0.0
2.5
5.0

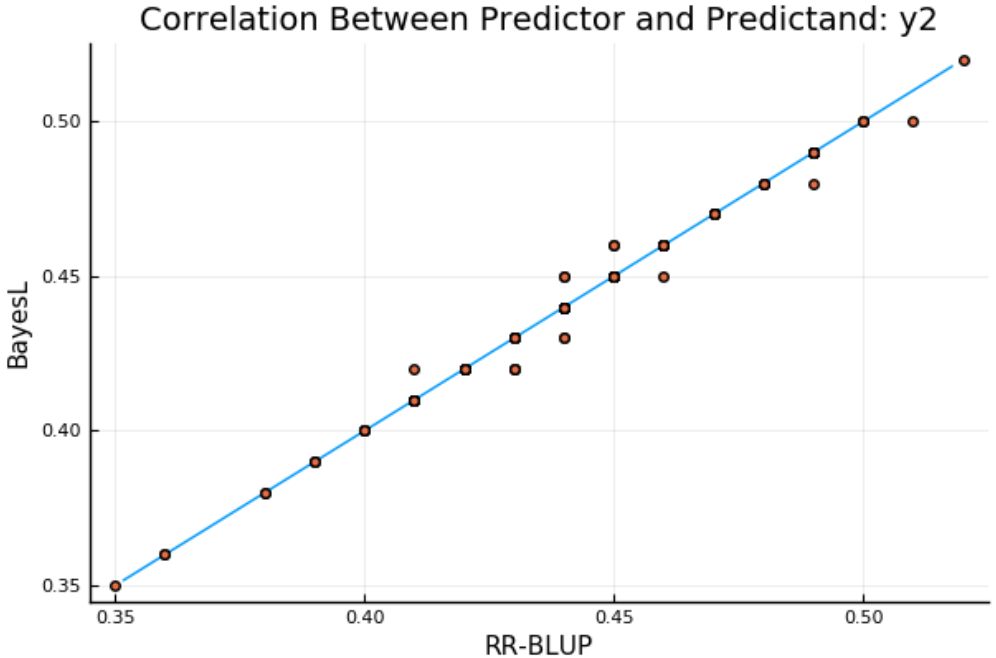

Standardized marker effects (1279) on yield 2: OLS vs MBL
t−stat MBL
t−stat OLS
1.0
0.5
0.0
−0.5
−1.0
−6
−3
0
3
6

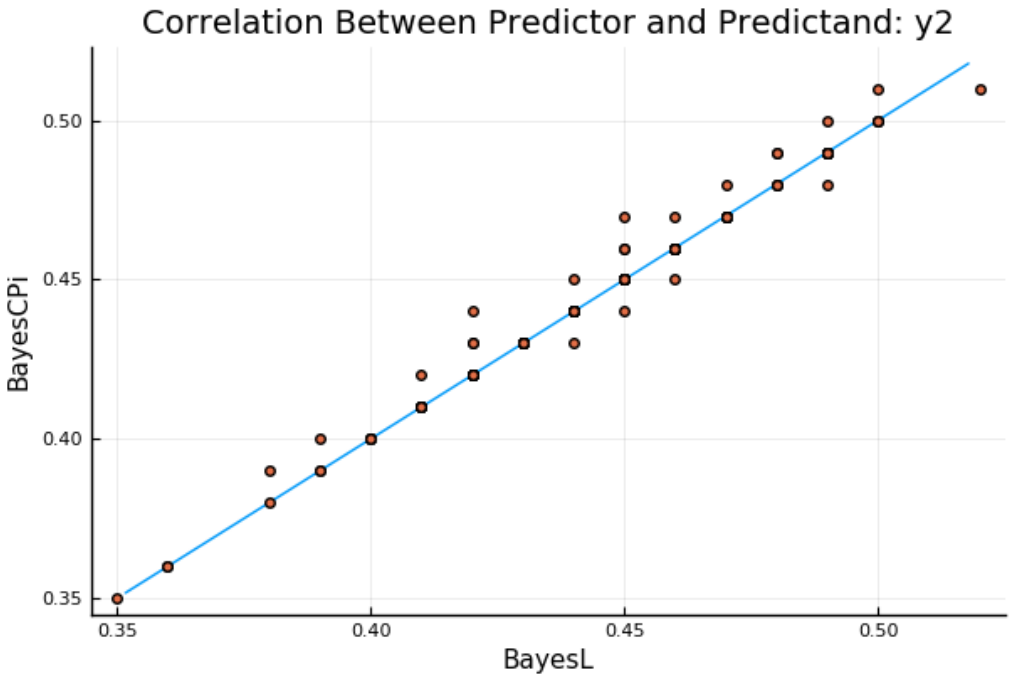

Standardized marker effects (1279) on yield 2: OLS vs BLUP
t−stat BLUP
t−stat OLS
1
0
−1
−6
−3
0
3
6

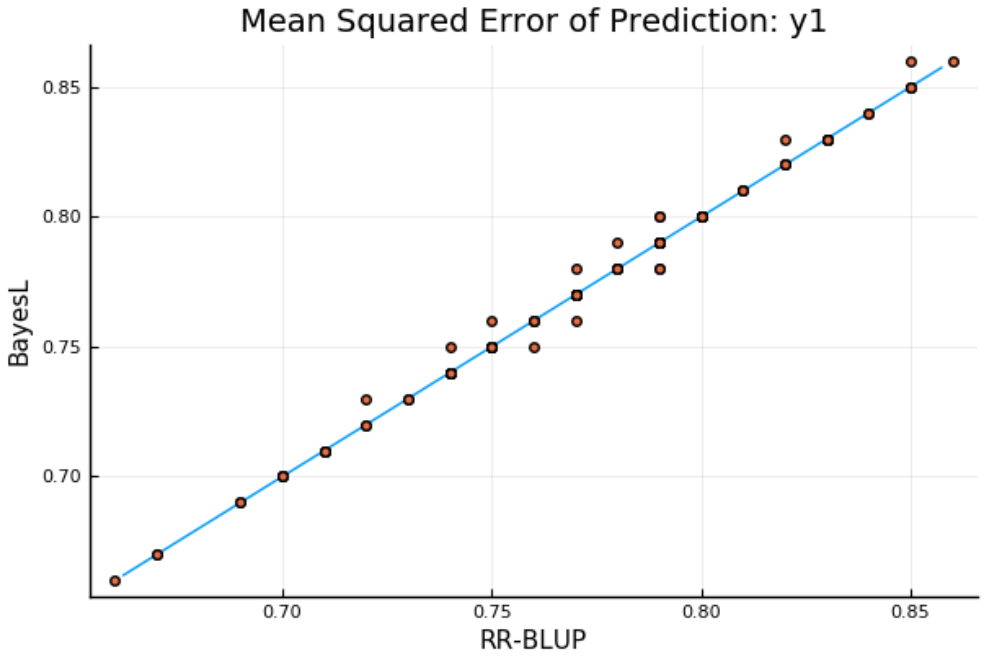

Mahalanobis squared distances (M)
(45 degree line in black))
BLASSO
GBLUP
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0

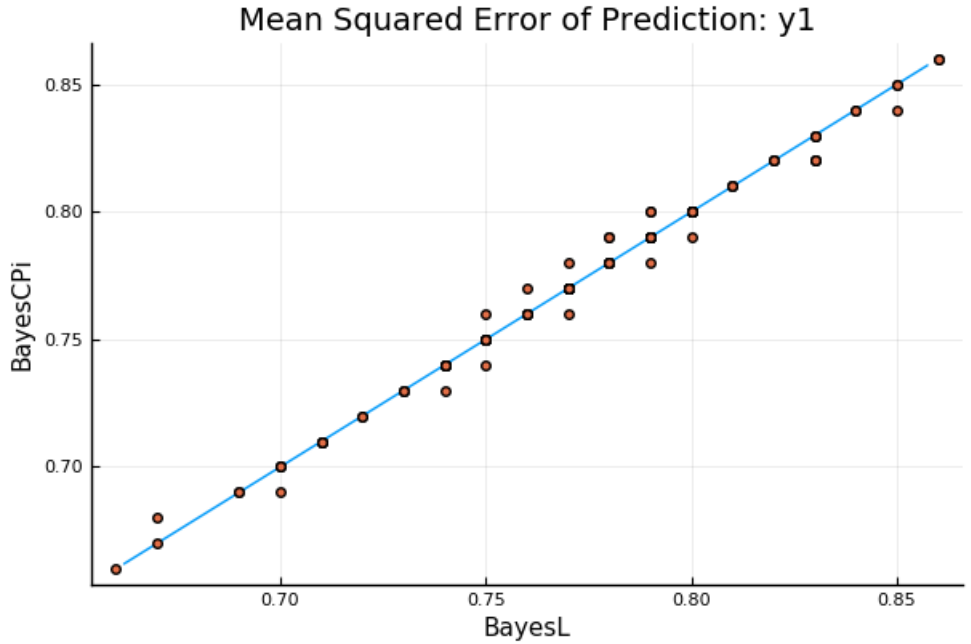

Correlation Between Predictor and Predictand: y1
BayesL
RR-BLUP
0.40
0.45
0.50
0.55

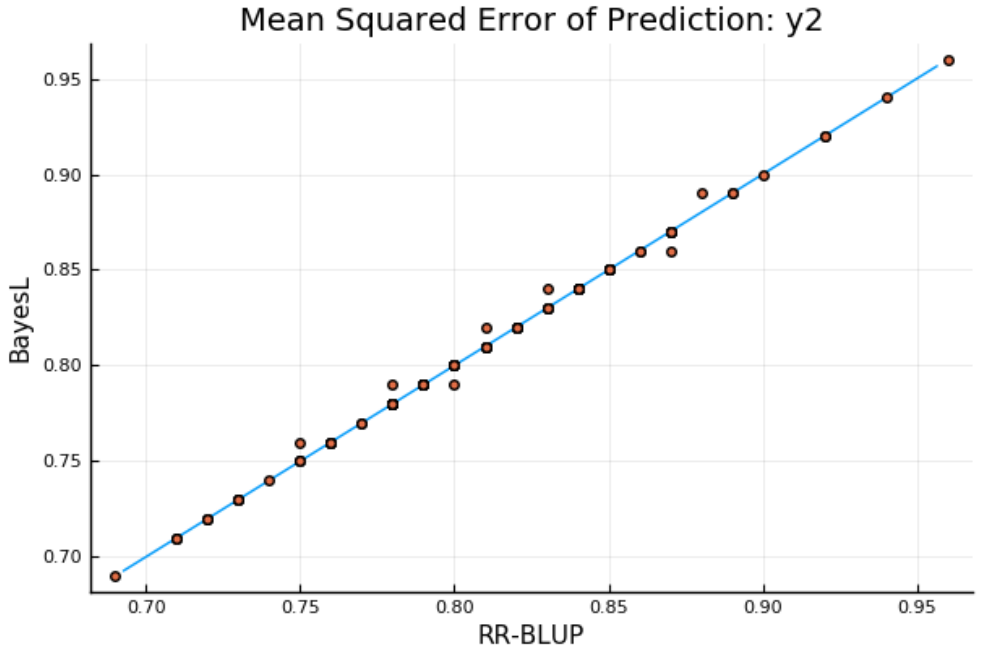

Correlation Between Predictor and Predictand: y1
BayesCPi
BayesL
0.40
0.45
0.50
0.55

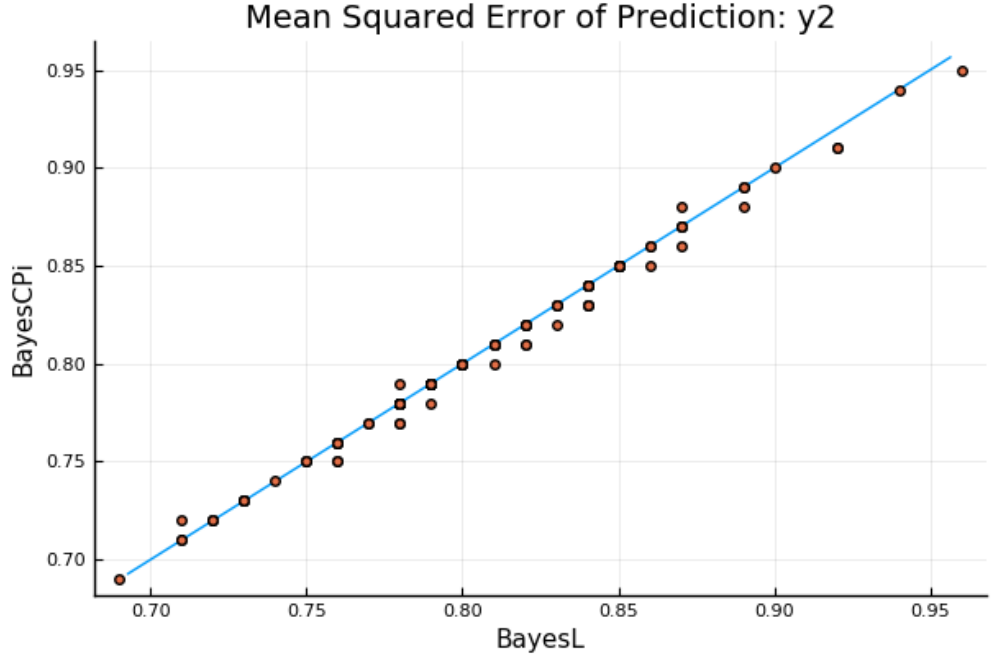

Correlation Between Predictor and Predictand: y2
BayesL
RR-BLUP
0.35
0.40
0.45
0.50


Correlation Between Predictor and Predictand: y2
BayesCPi
BayesL
0.35
0.40
0.45
0.50


Mean Squared Error of Prediction: y1
BayesL
RR-BLUP
0.70
0.75
0.80
0.85


Mean Squared Error of Prediction: y1
BayesCPi
BayesL
0.70
0.75
0.80
0.85


Mean Squared Error of Prediction: y2
BayesL
RR-BLUP
0.70
0.75
0.80
0.85
0.90
0.95


Mean Squared Error of Prediction: y2
BayesCPi
BayesL
0.70
0.75
0.80
0.85
0.90
0.95