A Multiple-Trait Bayesian Lasso for Genome-Enabled Analysis and Prediction of Complex Traits

Daniel Gianola*†‡§ and Rohan L. Fernando‡

*Department of Animal Sciences, and ‡Department of Dairy Science, University of Wisconsin-Madison, Wisconsin 53706, †Department of Animal Science, Iowa State University, Ames, Iowa 50011, and §Department of Plant Sciences, Technical University of Munich (TUM), TUM School of Life Sciences, Freising, 85354 Germany

ORCID IDs: 0000-0001-8217-2348 (D.G.); 0000-0001-5821-099X (R.L.F.)

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 trait’s 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 \( \Sigma \). 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 \( \Sigma \) 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 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 toward the origin than in GBLUP; conversely, “long” vectors are shrunk less. A predictive comparison was carried out as well in wheat, where the comparators of MBL were bivariate GBLUP and bivariate Bayes C—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 than 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 resource in the armamentarium for genome-enabled analysis and prediction of complex traits.

KEYWORDS Bayesian Lasso; complex traits; GWAS; multiple-trait; prediction; quantitative genetics; quantitative genomics

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., de los Campos et al. 2010; Yang et al. 2010; Lee 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 long been of great interest in plant and animal breeding, 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 has gradually become routine in the field. For example, Gao et al. (2018) described an application of a nine-variate model to data representing close to 7 million and 4 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 root 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; VanRaden 2007). Other procedures, such as Bayesian mixture models, are more involved, but extensions are available (Calus and Veerkamp 2011; Jia and Jannink 2006). Bayesian approaches are particularly interesting because they provide a framework for the joint estimation of the effects of multiple traits, allowing for the simultaneous consideration of genetic and environmental factors. However, Bayesian methods can be computationally intensive and require careful specification of prior distributions.

For instance, Gao et al. (2018a) used Bayesian model averaging, so posterior distributions are consistent 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 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 first application of BL in quantitative genetics was made by Yi and Xu (2008) in the context of quantitative trait locus (QTL) mapping, with subsequent applications reported in de los Campos et al. (2009), Legarra et al. (2011), Li et al. (2011) and Lehermeier et al. (2013).

It appears that no multiple-trait generalization of the BL has yet been reported. 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. The section Multi-Trait Regression Model describes MBL, including a MCMC 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.
Multi-Trait Regression Model

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

$$
y_i = \mu + \sum_{j=1}^{p} x_{ij} \beta_j + e_i; \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, p, \quad (1)
$$

where $y_i$ is a $T \times 1$ vector of responses observed in individual $i$, $\mu = \{\mu_1\}$ is the vector of trait means, and $x_{ij}$ is the genotype individual $i$ possesses at marker locus $j$. The residual vector $e_i$ ($T \times 1$) is assumed to follow the Gaussian distribution $e_i | R_0 \sim \mathcal{N}(0, R_0)$, where $R_0$ is a $T \times T$ covariance matrix. All $e_i$ vectors are assumed to be mutually independent and identically distributed.

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

$$
p\left(y_1, y_2, \ldots, y_N | \mu, \beta_1, \beta_2, \ldots, \beta_p, R_0\right)
$$

where

$$
P \left( y_i \right) = \frac{1}{\left|R_0\right|^2} \exp \left[ -\frac{1}{2} \sum_{i=1}^{N} \left( y_i - \mu - \sum_{j=1}^{p} x_{ij} \beta_j \right)^T R_0^{-1} \left( y_i - \mu - \sum_{j=1}^{p} x_{ij} \beta_j \right) \right].
$$

Bayesian prior assumptions

Parameters $\mu$ and $R_0$: The vector $\mu$ will be assigned a “flat” improper prior, and Jeffreys noninformative prior (e.g., Sorensen and Gianola 2002) will be adopted for $R_0$ so that their joint prior density is

$$
p(\mu, R_0) \propto \left| R_0 \right|^{-\left( \frac{T+1}{2} \right)}.
$$

Multivariate Laplace prior distribution 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 $\beta_j$ ($j = 1, 2, \ldots, p$), assumed mutually independent, a priori. Gómez et al. (1998) 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

$$
p(\beta_j | \Sigma) = \frac{T \Gamma\left(\frac{T}{2}\right)}{\left|\Sigma\right|^2 \pi^\frac{T}{2}} \exp\left(\frac{1}{2} \sqrt{\beta_j^T \Sigma^{-1} \beta_j}\right);
$$

where $\Sigma = \{\Sigma_t\}$ is a $T \times T$ positive-definite scale matrix. The variance–covariance matrix of MLAP is

$$
\text{Var}(\beta_j | \Sigma) = 4(T+1) \Sigma = \mathbf{B};
$$

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

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

The preceding is the density of a DE distribution with null mean, parameter $\sqrt{4\Sigma}$ and variance $\text{Var}(\beta) = 8\Sigma$. As mentioned earlier, Tibshirani (1996) and Park and Casella...
Gómez-Sánchez-Manzano et al. (2008) used the DE distribution as conditional (given $\Sigma$) 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. (1998) showed that such densities are elliptically contoured, and, thus, not DE. Appendix A and Supplemental Material, 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) $[\beta_j | \Sigma, \nu^2_j] = N_T(0, \nu^2_j \Sigma)$, and (2) $\nu^2_j \sim \text{Gamma}(\frac{\nu_0}{2}, \frac{1}{2})$ for $j = 1, 2, \ldots, p$; $N_T(0, \Sigma \nu^2_j)$ denotes a $T$–variate normal distribution with null mean and covariance matrix $\Sigma \nu^2_j$. The density of $\nu^2_j$ is

$$h(\nu^2_j) \propto (\nu^2_j)^{\frac{\nu_0}{2} - 1} \exp\left(-\frac{\nu^2_j}{2}\right).$$ (10)

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

$$\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T.$$ (11)

If independent and identical MLAP prior distributions are assigned to each of the subvectors, the joint prior density of all marker effects, given $\Sigma$, can be represented as

$$p(\beta | \Sigma) = \prod_{j=1}^{p} p(\beta_j | \Sigma) = \prod_{j=1}^{p} \int_0^\infty N_T(\beta_j | 0, \Sigma \nu^2_j) h(\nu^2_j) d\nu^2_j,$$ (12)

and the joint density of $\beta$ and $\nu^2 = [\nu^2_1, \nu^2_2, \ldots, \nu^2_p]^T$ is

$$p(\beta, \nu^2 | \Sigma) = \prod_{j=1}^{p} N_T(0, \Sigma \nu^2_j) h(\nu^2_j).$$ (13)

When individuals are sorted within traits (e.g., $T = 3$), note that $[\beta^T | \Sigma, \nu^2]$ is a $Tp$–dimensional normal distribution, with null mean vector and covariance matrix

$$\text{Var}[\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} | \Sigma, \nu^2] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix} = \Sigma \otimes D,$$ (14)

where $D = \text{diag}(\nu^2_1, \nu^2_2, \ldots, \nu^2_p)$ is a diagonal matrix. Hence,

$$p(\beta^T | \Sigma, \nu^2) \propto \exp\left[-\frac{1}{2} \beta^T (\Sigma^{-1} \otimes D^{-1}) \beta^T\right].$$ (15)

**Scale matrix $\Sigma$:** The scale matrix $\Sigma$ 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 $\Omega_\beta$ and $\nu_\beta$ degrees of freedom will be assigned as prior. The density is

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

**Joint posterior and fully conditional distributions**

The joint posterior distribution, including $\nu^2 = \{\nu^2_j\}$ from the scale-mixture of normals representation of the prior distribution of $\beta$, was assumed to take the form

$$p(\mu, \beta^*, R_0, \Sigma, \nu^2 | y^*, H) \propto p(y^* | \mu, \beta^*, R_0) p(R_0 | H)$$

$$\times p(\beta^* | \Sigma, \nu^2) p(\nu^2 | \Sigma | H),$$ (17)

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

**Parameters $\mu$ and $\beta^*$ given ELSE:** From (17,), and using representations (4) and (15), the fully conditional posterior distribution of $\mu$ and $\beta^*$ has density

$$p(\mu, \beta^* | \text{ELSE}) \propto p(y^* | \mu, \beta^*, R_0) p(\beta^* | \Sigma, \nu^2)$$

$$\times \exp\left[-\frac{1}{2} (y^* - \bar{y}(\mu, \beta^*))' R^{-1} (y^* - \bar{y}(\mu, \beta^*))\right]$$

$$\times \exp\left[-\frac{1}{2} \beta^* (\Sigma^{-1} \otimes D^{-1}) \beta^*\right].$$ (18)

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

$$\begin{bmatrix} \mu \\ \beta^* \end{bmatrix} = \begin{bmatrix} R_0^{-1} N & R_0^{-1} \otimes 1_N' X \\ R_0^{-1} \otimes X' 1_N & R_0^{-1} \otimes X' X + \Sigma^{-1} \otimes D^{-1} \end{bmatrix}^{-1} \times \begin{bmatrix} R_0^{-1} \otimes 1_N' y^* \\ (R_0^{-1} \otimes X)' y^* \end{bmatrix},$$ (19)

and the variance–covariance matrix is
A more explicit representation is presented in Appendix B for the case \( T = 3 \).

**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 \( \Sigma, \nu^2 \), and \( R_0 \), is multivariate normal, all conditionals and linear combinations thereof are normal as well. In particular \((T = 3)\),

\[
E(\mu_i | \text{ELSE}) = \frac{1}{r_t N} \left[ 1_N^t \sum_{t' = 1}^3 r_{t'}^t (y_{t'} - X \beta t) - N\sum_{t' \neq i} r_{t'}^t \mu_{t'} \right];
\]

\( i = 1, 2, 3, \)

and

\[
\text{Var}(\mu_i | \text{ELSE}) = \frac{1}{r_t^2 N}; \quad t = 1, 2, 3, \quad (21)
\]

Likewise

\[
E(\beta_j^i | \text{ELSE}) = \frac{1}{N} \left[ X^t \sum_{t = 1}^3 r_{t'}^t (y_j^t - 1N \mu_j) - \sum_{t' \neq j} (r_{t'}^t X^t X + D^{-1}) \beta_j^t \right];
\]

\( i = 1, 2, 3, \)

and

\[
\text{Var}(\beta_j^i | \text{ELSE}) = \left( \frac{r_{i}^t X^t X + D^{-1}}{\Sigma_{tt}} \right)^{-1}; \quad t = 1, 2, 3. \quad (22)
\]

**Fully conditional distributions of \( R_0 \) and \( \Sigma \):** From (17), using (2) and (6)

\[
p(R_0 | \text{ELSE}) \propto |R_0| \left( \frac{N + T + 1}{2} \right) \exp \left\{ -\frac{1}{2} \text{tr}(R_0^{-1} S_e) \right\},
\]

so \( |R_0| \text{ELSE} \) is an \( IW \) distribution with \( N + T \) degrees of freedom and scale matrix \( S_e \). In \( IW \), the kernel of the density is often written as \( \exp\left\{ -\frac{1}{2} \text{tr}(R_0^{-1} (N + T) S_e) \right\} \), where \( S_e = S_e / (N + T) \).

Recall that

\[
p(\beta^j | \Sigma, \nu^2) = \prod_{j=1}^p N_T(\beta_j | 0, \Sigma \nu_j^2), \quad (26)
\]

so, from (17)

\[
p(\Sigma | \text{ELSE}) \propto p(\beta^j | \Sigma, \nu^2)p(\Sigma | H)
\]

\[
\propto \prod_{j=1}^p \frac{1}{|\nu_j^2|^2} \exp \left\{ -\frac{1}{2} \beta_j^t (\Sigma^{-1} \beta_j) \right\} - \frac{(T + \nu_B + 1)}{2} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma^{-1} \Omega_{\beta}) \right\} \exp \left\{ -\frac{1}{2} \text{tr}(\Sigma^{-1} S_\beta) \right\}, \quad (27)
\]

where

\[
S_\beta = \sum_{j=1}^p \left( \frac{\beta_j \beta_j^t}{\nu_j^2} \right) + \Omega_{\beta} \quad (28)
\]

is a \( T \times T \) matrix. Hence the conditional posterior distribution of \( \Sigma \) is \( IW(p + T + \nu_B, S_\beta) \). The kernel of the density of \( \Sigma \) is often represented as \( \exp\left\{ -\frac{1}{2} \text{tr}(\Sigma^{-1} (p + T + \nu_B) S_\beta) \right\} \), where \( S_\beta = S_\beta / (p + T + \nu_B) \).

**Fully conditional distribution of \( \nu^2 \):** From (17), and using (13),

\[
p(\nu^2 | \text{ELSE}) \propto p(\beta^j | \Sigma, \nu^2)p(\nu^2)
\]

\[
\propto \prod_{j=1}^p N_T(\beta_j | 0, \Sigma \nu_j^2) h_1(\nu_j^2) \exp \left\{ -\frac{\beta_j \Sigma^{-1} \beta_j}{2 \nu_j^2} \right\} \frac{\nu_j^{\nu_j + 1}}{\Gamma(\nu_j + 1)} \exp \left\{ -\frac{\nu_j^2}{8} \right\}
\]

\[
\propto \prod_{j=1}^p \left( \nu_j^2 \right)^{-\frac{\nu_j}{2}} \exp \left\{ -\frac{\beta_j \Sigma^{-1} \beta_j + \nu_j^2}{2 \nu_j^2} \right\}. \quad (29)
\]

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.

**MCMC algorithm**

Starting values for \( R_0 \) and \( \Sigma \) can be obtained “externally” from some estimates of \( R_0 \) and \( B \) (the \( T \times T \) matrix of
variances and covariances of marker effects) calculated with standard methods such as maximum likelihood. Recall that \( \Sigma = B/[4(T + 1)] \).

Sample each \( \nu_j^2 (j = 1, 2, \ldots, p) \) 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 \( \nu_j^{2[t+1]} = 2/\nu_j^t \) as a new state. Otherwise, set \( \nu_j^{2[t+1]} = \nu_j^t ; j = 1, 2, \ldots, p \).

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

Sample \( R_0 \) from \( IW(N + T, S_e) \) and \( \Sigma \) from \( IW(p + T + v_p, S_p) \).

**Remarks**

Appendix E shows that the degree of shrinkage of marker effects results from a joint action between \( \Sigma \) and the strength of marker effects. A vector of effects of a marker with a short Mahalanobis distance away from \( 0 \) is more strongly shrunk toward 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 “pseudoselection” 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.

**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.

**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.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. 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 C\(_p\). This wheat data set has been studied extensively, e.g., by Crossa et al. (2010), Gianola et al. (2011, 2016), and Long et al. (2011). 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 the 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 breeding of dairy cattle, 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 (\textit{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 vs. 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). Supplemental material available at figshare: https://doi.org/10.25386/genetics.10689380.

Bivariate Analysis of Wheat Yield: MBL vs. GBLUP

\textbf{Genomic BLUP and Bayesian BLUP}

The bivariate model was

\[
\begin{bmatrix}
y_1 \\
y_2
\end{bmatrix} = \begin{bmatrix} 1 & \mu_1 \\ 1 & \mu_2 \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix},
\]

where \( y_1(y_2) \) is the vector of grain yields in environment 1(2) of the 599 inbred lines; \( \mu_1 \) and \( \mu_2 \) are the trait means in the two environments, and \( \mathbf{1} \) is a \( 599 \times 1 \) incidence vector of ones; \( g_1 \) and \( g_2 \) are the “additive genomic values” of the lines, and \( e_1 \) and \( e_2 \) are model residuals. In GBLUP (VanRaden 2008) the genetic signals captured by markers are represented as \( g_1 = X \beta_1 \) and \( g_2 = X \beta_2 \), where \( X \) is a \( 599 \times 1279 \) centered and scaled matrix of genotype codes, and \( \beta_1(\beta_2) \) contains the marker allele substitution effects on trait 1(2). The residual distribution was...
where, as before, $R_0$ is the $2 \times 2$ 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 $\beta_1 \sim N(0, I \sigma^2_{\beta_1})$, $\beta_2 \sim N(0, I \sigma^2_{\beta_2})$ and $\text{Cov}(\beta_1, \beta_2) = I \sigma_{\beta_1 \beta_2}$, so

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

(32)

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

$$
\begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} \sim N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, G_0 \otimes G\right),
$$

(33)

where $G_0 = pB = \begin{bmatrix}
\sigma^2_g_1 \\
\sigma_{g_1 g_2} \\
\sigma_{g_1 g_2} \\
\sigma^2_g_2
\end{bmatrix}$

(34)

is a between-trait variance–covariance matrix of the additive genomic values (here, e.g., $\sigma^2_g_2 = p \sigma^2_{\beta_2}$) and $G = XX'/p$ is a genomic-relationship matrix describing genome-based similarities among the 599 lines. The preceding assumptions induce the marginal distribution

$$
\begin{bmatrix}
y_1 \\
y_2
\end{bmatrix} \sim N\left(\begin{bmatrix}1 \mu_1 \\ 1 \mu_2\end{bmatrix}, V = G_0 \otimes G + R_0 \otimes I\right),
$$

(35)

where $V$ is the phenotypic covariance matrix. The bivariate best linear unbiased predictor of $g_1$ and $g_2$ (Henderson 1975) is:

$$
\begin{bmatrix}
\hat{g}_1 \\
\hat{g}_2
\end{bmatrix} = (G_0 \otimes G)V^{-1}\left(\begin{bmatrix}y_1 - 1 \mu_1 \\ y_2 - 1 \mu_2\end{bmatrix}\right),
$$

(36)

where

$$
\begin{bmatrix}
\hat{\mu}_1 \\
\hat{\mu}_2
\end{bmatrix} = \left(\begin{bmatrix}1' & 0' \\ 0 & 1\end{bmatrix}V^{-1}\left[\begin{bmatrix}1 \\ 0\end{bmatrix} \ 0 \ 1\end{bmatrix}\right]
\right)^{-1}\left(\begin{bmatrix}1' \\ 0\end{bmatrix}V^{-1}\left[\begin{bmatrix}y_1 \\ y_2\end{bmatrix}\right]\right),
$$

(37)

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

BLUP and GLS require knowledge of $G_0$ and $R_0$, 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 $\text{Cov}(X, Y) = [\text{Var}(X + Y) - \text{Var}(X) - \text{Var}(Y)]/2$. The resulting estimates of $G_0$ and $R_0$ were inside their corresponding parameter spaces. An estimate of $B$ was obtained by applying relationship (34) to the estimated $G_0$.

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

Figure 3 Bivariate GBLUP vs. bivariate Bayesian LASSO (posterior mean) estimates of marker effects on wheat grain yield.
marker effects \((\beta)\) given predictions of signal \((g)\). If \(\beta\) and \(g\) have a joint normal distribution, under (30) one has

\[
E\left(\begin{bmatrix} \beta_1 \\ \beta_2 \\ g_1 \\ g_2 \end{bmatrix} \bigg| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right) = (BG_0^{-1} \otimes X'G^{-1}) \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}.
\]

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

\[
\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = E\left(\begin{bmatrix} \beta_1 \\ \beta_2 \\ g_1 \\ g_2 \end{bmatrix} \bigg| \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right) = (BG_0^{-1} \otimes X'G^{-1}) \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = (BG_0^{-1} \otimes X'G^{-1})V^{-1}\left(\begin{bmatrix} y_1 - \hat{\mu}_1 \\ y_2 - \hat{\mu}_2 \end{bmatrix}\right),
\]

with \(\hat{\beta}_i = E(\beta_i|y_1, y_2), i = 1, 2\). After lengthy algebra, and using Henderson (1975), the prediction error variance–covariance matrix of the BLUP of marker effects is given by

\[
\text{Var}\left(\begin{bmatrix} \beta_1 - \beta_1 \\ \beta_2 - \beta_2 \end{bmatrix}\right) = (B \otimes I_p) - (B \otimes X')\left[I_{2n} + V^{-1} - \frac{V^{-1}11'V^{-1}}{1'V^{-1}1}\right](B \otimes X').
\]

(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 \(B\) and \(R_0\) 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, $G_0$ and $R_0$ were treated as unknown parameters, with the uncertainty about their values accounted for.

**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 $\Sigma$ in turn, $\Sigma$ was given a two-dimensional (2D) inverse Wishart distribution on $v_\mu = 4$ degrees of freedom and with scale matrix $\Omega_\mu = v_\mu B/12 = B/3$. The residual variance–covariance matrix $R_0$ was assigned the 2D Jeffreys improper prior in (6).

The MCMC scheme employed the scale mixture of normal 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 $m_1, m_2, R_0, \Sigma$, 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. Figures S4–S13 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. Figures S14 and S15 depict post burn-in trace plots for the elements of $R_0$ and $\Sigma$, respectively. The six chains “joined” eventually, and sample values fluctuated thereafter within what seemed to be stationary distributions. To assess convergence further, a test suggested by Geweke (1992) was applied to the combined 12,000 samples from the posterior distributions of $\mu_1, r_{12}$.
(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.

Figure 1 and Figure 2 display estimated posterior densities of \( r_{\varepsilon 12} \) and \( r_{\beta 12} \). Mixing for \( r_{\beta 12} \) was poorer than for \( r_{\varepsilon 12} \), 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 bivariate BLUP of marker effects.

Location parameters mixed well. For example, the average effective sample size of \( \mu_2 \) over the six 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, the 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.

**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 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 \( t-statistic \) from the GWAS-type analysis with the standardized BLUP and MBL effects is provided in Figure 5. As expected, shrinkage toward null-mean distributions (bivariate Gaussian in BLUP and bivariate Laplace in MBL) made \( t-statistic \) 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 (\( m^2 \)) 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 \( m^2_{Blup,j} = \beta_{Blup,j}^T \Sigma_{Blup}^{-1} \beta_{Blup,j} \), and for MBL it was \( m^2_{MBL,j} = \beta_{MBL,j}^T (12 \Sigma)^{-1} \beta_{MBL,j} \), where \( \beta_j \) are effect estimates for marker \( j \), and \( \Sigma \) is the estimated posterior expectation of \( \Sigma \). 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 \( t-statistic \).

**Predictive Comparison between MBL, MTGBLUP, and MT-BayesC\( \pi \): Wheat**

Bivariate Bayesian GBLUP and BayesC\( \pi \) 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); BayesC\( \pi \) consisted of a bivariate mixture in which each of the 1279 markers was allowed to fall, a priori.
into one of four disjoint classes: \((0, 0), (0, 1), (1, 0), \) and \((1, 1),\)
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 \(\text{Dirichlet}(1, 1, 1, 1)\) 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 1000 draws. The analyses were run using the JWAS package written in the JULIA language (Cheng et al. 2018b).

Figure 7 and Figure 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 \(100 (X, Y)\) 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.

**Predictive Comparison between MBL vs. MTGBLUP and MBL vs. Single Trait Bayesian LASSO: Pinus**

Figure 9 and Figure 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 instances in which a multiple-trait specification clearly constitutes a better prediction machine.
Conclusion

Our study is possibly the first report in the quantitative genetics literature of a 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 DE prior of the BL when applied to a single trait. The implementation of MBL requires MCMC 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, in view of the large variability of performance with respect to data structure for any given prediction machine, it is largely futile to categorize methods in terms of expected predictive performance using broad criteria (Morota and Gianola 2014; Gianola and Rosa 2015; Momen et al. 2018; Montesinos-López et al. 2018 a,b, 2019a,b; Azodi et al. 2019).

MBL is expected to shrink more strongly toward 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 $\beta_j^\top \Sigma^{-1} \beta_j$. Thus, multivariate
Bayesian pseudosparsity is induced by MBL to an extent depending on the heterogeneity of $\beta_j^2$ over markers.

We note, in passing, that the term $P_p j = \frac{1}{\sqrt{\sum_j \beta_j^2}}$ given in (66) in Appendix E is the counterpart of $P_g = \frac{1}{\sqrt{\sum_g \beta_g^2}}$, 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 $\beta_g^2$ is the group regression coefficient for trait $t$. The latter penalty assigns the same weight to these regressions over traits, contrary to MBL, where weights and coweights are driven by $\Sigma^{-1}$. 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. (2018a). 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. Neither 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, $g$, or signal as referred to in our paper. In MBL, $g$ and marker effects can be inferred from their posterior mean or from a modal approximation (MAP-MBL) that does not involve MCMC that 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 $\Sigma = G_0/(12p)$, and

**Figure 10** Predictive correlation for rust bin and gall volume in pine trees: bivariate Bayesian LASSO (Bayes L) vs. bivariate Bayesian GBLUP (RR-BLUP).
starting values for the iteration were calculated using BLUP estimates of marker effects. MAP with $T = 2$ were iterated for 500 rounds. Figure S16 shows that, at iteration 500, the metric used for monitoring convergence had stabilized at the third decimal place, but, for our purposes, iteration could have stopped after 200 rounds. 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, the MAP approach gave markedly different results, producing a stronger shrinkage to 0 of small-effect markers and, thus, an effectively sparser model. Figure S18 gives a comparison of the fitted genomic values, i.e., $g_1 = X\beta_1^*$ and $g_2 = X\beta_2^*$ for the two traits. GBLUP and MBL estimates were closely aligned and fit 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 BayesC$\pi$, 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 et al. 2012; de los Campos et al. 2013; Momen et al. 2018; Bellot et al. 2018; Montesinos-López et al. 2018a,b, 2019a,b), even when nonparametric or deep learning techniques are brought into the comparisons. For this reason, we refrain from making any sweeping claim about the superiority (inferiority) of MBL over any other competing multiple-trait Bayesian regression method. If the situation is
such that multiple-trait vectors of effects are fairly homogeneous over makers, it is to be expected that most methods will have a similar performance. On the other hand, if there is underlying heterogeneity of vectors of effects, possibly reflecting “structural sparsity” at the QTL level, it is quite likely that MBL and multiple-trait Bayesian variable selection methods (e.g., Cheng et al. 2018a) will outperform MTGBLUP or a multiple-trait version of Bayes A. Unfortunately, is it difficult to anticipate a priori which method will deliver the best performance, given the limited knowledge of the biological architecture of complex traits, the strong influence of the data structure, and the variability of $X$ matrices over data sets, in dimension and content.

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 implementation of the quantitative genetics literature of a multiple-trait LASSO, 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 qualification. 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).

Acknowledgments

Chris-Carolin Schön 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 D.G. was provided by the Deutsche Forschungsgemeinschaft (grant No. SCH0 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, the Technical University of Munich (TUM) Institute for Advanced Study, TUM School of Life Sciences, and the Institut Pasteur de Montevideo, Uruguay, are thanked for providing office space, library and computing resources, and logistic support.

Literature Cited

Azodi, C., E. Bolger, A. McCarren, M. Roantree, G. de los Campos et al., 2019 Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 (Bethesda) 9: 3691–3702. https://doi.org/10.1534/g3.119.400498

Bellot, P., G. de Los Campos, and M. Pérez-Enciso, 2018 Can deep learning improve genomic prediction of complex traits? Genetics 210: 809–819. https://doi.org/10.1534/genetics.118.301298

Calus, M. P. L., and R. F. Veerkamp, 2011 Accuracy of multi-trait genomic selection using different methods. Genet. Sel. Evol. 43: 26. https://doi.org/10.1186/1297-9686-43-26

Cheng, H., K. Kizilkaya, J. Zeng, and D. Garrick, 2018a Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics 209: 89–103. https://doi.org/10.1534/genetics.118.300650

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

Clossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño et al., 2010 Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713–724. https://doi.org/10.1534/genetics.110.118521

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra et al., 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385. https://doi.org/10.1534/genetics.109.101501

de los Campos, G., D. Gianola, and D. A. B. Allison, 2010 Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. 11: 880–886. https://doi.org/10.1038/nrg298

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193: 327–345. https://doi.org/10.1534/genetics.112.143313

Falconer, D. S., 1952 The problem of environment and selection. Am. Nat. 86: 293–298. https://doi.org/10.1086/281736

Fernando, R., A. Toosi, A. Wolfe, D. Garrick, and J. Dekkers, 2017 Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. J. Agric. Biol. Environ. Stat. 22: 172–193. https://doi.org/10.1007/s12295-017-0277-6

Galesloot, T. E., K. van Steen, L. A. L. M. Kiemeney, L. L. Janss, and S. H. Vermeulen, 2014 A comparison of multivariate genome-wide association methods. PLoS One 9: e95923. https://doi.org/10.1371/journal.pone.0095923

Gao, H., P. Madsen, J. Pöösö, G. P. Aamand, M. Lidauer et al., 2018 Short communication: Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population. J. Dairy Sci. 101: 11159–11164. https://doi.org/10.3168/jds.2018-15123

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

Gelman, A., and D. B. Rubin, 1992 Inference from iterative simulation using multiple sequences (with discussion). Stat. Sci. 7: 457–472. https://doi.org/10.1214/ss/1177011136

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

Geweke, J., 1992 Evaluating the accuracy of sampling-based approaches to calculating posterior moments, pp. 169–193 in Bayesian Statistics 4, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, Clarendon Press, Oxford.

Gianola, D., 2013 Priors in whole genome regression: the Bayesian alphabet returns. Genetics 194: 573–596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., and G. J. M. Rosa, 2015 One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biol. 3: 19–56. https://doi.org/10.1146/annurev-animal-022114-110733

Gianola, D., M. Pérez-Enciso, and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. Genetics 63: 347–365.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando, 2009 Additive genetic variability and the Bayesian
Multiple-Trait Bayesian Lasso 323

analysis of complex traits using a Bayesian mixture model. PLoS Genet. 11: e1004969. https://doi.org/10.1371/journal.pgen.1004969

Park, T., and G. Casella, 2008 The Bayesian LASSO. J. Am. Stat. Assoc. 103: 681–686. https://doi.org/10.1198/016214508000000337

Pérez, P., and G. de los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483–495. https://doi.org/10.1534/genetics.114.164442

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

Schaid, D. J., C. Wenan, and N. B. Larson, 2018 From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. 19: 491–504. https://doi.org/10.1038/s41576-018-0016-z

Singh, G., G. S. Bhullar, and K. S. Gill, 1986 Genetic control of grain yield and its related traits in bread wheat. Theor. Appl. Genet. 72: 536–540. https://doi.org/10.1007/BF00289537

Sleper, D. A., and M. Poehlman, 2006 Breeding Field Crops, Ed. 5, Wiley-Blackwell, Hoboken, NJ.

Smith, H. F., 1936 A discriminant function for plant selection. Ann. Eugen. 7: 240–250. https://doi.org/10.1111/j.1469-1809.1936.tb02143.x

Sorensen, D., and D. Gianola, 2002 Likelihood, Bayesian, and MCMC methods in quantitative genetics, Springer, New York. https://doi.org/10.1007/b98952

Strandén, I., and D. J. Garrick, 2009 Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92: 2971–2975. https://doi.org/10.3168/jds.2008-1929

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

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

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Van Tassell, C. P., and L. D. Van Vleck, 1996 Multiple-trait Gibbs sampler for animal models: flexible programs for Bayesian and likelihood-based (co)variance component inference. J. Anim. Sci. 74: 2586–2597. https://doi.org/10.2527/1996.74112586x

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. Am. J. Hum. Genet. 90: 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy et al., 2017 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. 101: 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Walsh, B., and M. Lynch, 2018 Evolution and selection of quantitative traits, Oxford University Press, New York. https://doi.org/10.1093/oso/9780198830870.001.0001

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders et al., 2010 Common SNP’s explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569. https://doi.org/10.1038/ng.608

Yi, N., and S. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045–1055. https://doi.org/10.1534/genetics.107.085589

Yuan, M., and Y. Lin, 2006 Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B 68: 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Yuan, M., A. Ekici, Z. Lu, and R. Monteiro, 2007 Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. B 69: 329–346. https://doi.org/10.1111/j.1467-9868.2007.00591.x

Communicating editor: M. Sillanpää
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:

\[
\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \Sigma_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}; \Sigma_3 = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}.
\]

(40)

Using (7), the density under \(\Sigma_1\) is

\[
p(\beta_1, \beta_2 | \Sigma_1) = \frac{1}{8\pi} \exp \left( -\frac{1}{2} \sqrt{\beta_1^2 + \beta_2^2} \right).
\]

(41)

The covariance matrix here, \(B_1 = 12\Sigma_1\), is diagonal, so the random variables are uncorrelated but not independent, since (41) cannot be written as the product of two marginal densities. Under \(\Sigma_2\) and \(\Sigma_3\), the densities are

\[
p(\beta_1, \beta_2 | \Sigma_2) = \frac{5\sqrt{6}}{96\pi} \exp \left( -\frac{1}{2} \sqrt{\frac{25}{24} \beta_1^2 - \frac{5}{12} \beta_1 \beta_2 + \frac{25}{24} \beta_2^2} \right),
\]

(42)

and

\[
p(\beta_1, \beta_2 | \Sigma_3) = \frac{5}{24\pi} \exp \left( -\frac{1}{2} \sqrt{\frac{25}{9} \beta_1^2 + \frac{40}{9} \beta_1 \beta_2 + \frac{25}{9} \beta_2^2} \right).
\]

(43)

Five bivariate Laplace densities are shown in Figure S1: (a) gives the density of the distribution of the two uncorrelated bivariate Laplace random variables (\(\Sigma_1\)), and (b) and (c) show the positively (i.e., with \(\Sigma_2\)) and negatively (with \(\Sigma_3\)) 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) display bivariate Laplace densities of distributions with non-null means.

13.1.2 Conditional distributions
Dropping subscript \(j\) denoting a specific marker, partitions the \(T \times 1\) vector of effects into \(\beta' = (\beta_1', \beta_2')\), where the subvectors have orders \(T_1\) and \(T_2\), respectively; recall that \(T\) is the number of traits. Correspondingly, put

\[
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.
\]

(44)

According to J. M. Marín (personal communication), the conditional distribution \(\beta_2 | \beta_1\) has density

\[
p(\beta_2 | \beta_1, \mu_{2|1}, \Sigma_{2|1}) = \frac{\Gamma \left( \frac{T - T_1}{2} \right)}{|\Sigma_{22}|^{\frac{T - T_1}{2}}} \int_0^\infty \exp \left\{ -\frac{1}{2} \sqrt{t + q_1} \right\} dt \times \exp \left\{ -\frac{1}{2} \sqrt{\left( \beta_2 - \mu_{2|1} \right)' \Sigma_{2|1}^{-1} \left( \beta_2 - \mu_{2|1} \right) + q_1} \right\},
\]

(45)

where \(\mu_{2|1} = \Sigma_{22} \Sigma_{11}^{-1} \mu_{1|1}, \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}\) and \(q_1 = \beta_1' \Sigma_{11}^{-1} \beta_1\). Similar to multivariate normal distribution, the conditional expectation is linear on the conditioning variable, and \(\Sigma_{2|1}\) does not involve \(\beta\).

13.1.3 Simulation of a multivariate Laplace distribution
Gómez et al. (1998) showed that \(S\) independent draws from a MLAP distribution with a null mean vector can be made as

\[
\beta_i = r_i C' u_i; \quad i = 1, 2, \ldots, S,
\]

(46)
where \( C' \) results from the Cholesky decomposition \( \Sigma = C'C \), \( u \) 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 \( u \) can be simulated by effecting \( T \) independent draws \( (x_i; \; i = 1, 2, \ldots, T) \) from a \( N(0, 1) \) distribution, and then forming the \( r \)th element of \( u \) as
\[
u_r = x_t \sqrt{\sum_{i=1}^{T} x_i^2}, \quad t = 1, 2, \ldots, T.
\]
Marginal distributions for the three bivariate Laplace distributions with scale matrices \( \Sigma_1, \Sigma_2, \) and \( \Sigma_3 \) given above were estimated by sampling \( S = 300,000 \) independent realizations; Equation 46 was employed. Using the samples, zero-mean DE and normal distributions with the same variances were estimated by sampling \( S = 300,000 \). 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. (1998): marginals from MLAP distributions are elliptically contoured and not DE.

### 13.2 Appendix B: Mean vector of location parameters given ELSE

Consider (Equation 19). For \( T = 3 \), let
\[
R_0^{-1} = \begin{bmatrix}
\rho_{11} & \rho_{12} & \rho_{13} \\
\rho_{21} & \rho_{22} & \rho_{23} \\
\rho_{31} & \rho_{32} & \rho_{33}
\end{bmatrix}
\quad \text{and} \quad
\Sigma^{-1} = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{31} & \Sigma_{32} & \Sigma_{33}
\end{bmatrix}
\]

Expansion of the Kronecker products in (Equation 19) produces the system
\[
\begin{bmatrix}
\rho_{11} & \rho_{12} & \rho_{13} & \rho_{11} \mathbf{X} & \rho_{12} \mathbf{X} & \rho_{13} \mathbf{X} \\
\rho_{21} & \rho_{22} & \rho_{23} & \rho_{21} \mathbf{X} & \rho_{22} \mathbf{X} & \rho_{23} \mathbf{X} \\
\rho_{31} & \rho_{32} & \rho_{33} & \rho_{31} \mathbf{X} & \rho_{32} \mathbf{X} & \rho_{33} \mathbf{X}
\end{bmatrix}
\begin{bmatrix}
\boldsymbol{P}_1 \\
\boldsymbol{P}_2 \\
\boldsymbol{P}_3
\end{bmatrix}
= \begin{bmatrix}
1_N (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
1_N (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
1_N (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*) \\
\mathbf{X} (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
\mathbf{X} (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
\mathbf{X} (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*)
\end{bmatrix}
\]
\[
= \begin{bmatrix}
1_N (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
1_N (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
1_N (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*) \\
\mathbf{X} (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
\mathbf{X} (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
\mathbf{X} (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*)
\end{bmatrix}
\]

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 \( Q = \beta' \Sigma^{-1} \beta \) take values \( \frac{1}{4}, 1, 4, \) 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 Figure S3. The distributions are skewed, and, as \( Q \) increases, the density becomes flatter.

Let \( S_1(y; \sigma) \) be the Lévy density of a positive random variable \( Y \) having a positive stable distribution with parameter \( \sigma \) (Samorödnitsky and Taqqu 2000). From Gómez et al. (1998) and Gómez-Sánchez-Manzano et al. (2008), the Lévy density is
\[
S_1(y; \sigma) = \frac{(\sigma^2 y^2/4)}{(\Gamma(3/2))} y^{-3/2-1} \exp \left( -\frac{\sigma}{4y} \right),
\]
(49)
which is that of an inverse Gamma (IG) distribution with parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{g}{4}$. Consider now the transformation (Gómez et al. 1998) $w_j = 2/y_j^2$ so using (29)

$$p(w_j|ELSE) \propto \left(\frac{2}{w_j}\right)^{-\frac{1}{2}} \exp\left[-\frac{Q_j + \frac{1}{4w_j}}{w_j}\right] \frac{2}{w_j^2}$$

$$\propto \sqrt{w_j}^{-1} \exp\left[-\frac{w_jQ_j}{4}\right] \exp\left[-\frac{1}{4w_j}\right], \quad (50)$$

Consider a Metropolis-Hastings ratio $R$ using (49) with $\alpha = 1$ 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

$$R = \frac{\sqrt{w_j}^{-1} \exp\left[-\frac{Q_j}{4}\right] \exp\left[-\frac{1}{4w_j}\right] \times w_j^{-\frac{1}{2}} \exp\left(-\frac{1}{4w_j}\right)}{\sqrt{y_j}^{-1} \exp\left[-\frac{Q_j(y_j)}{4}\right] \exp\left(-\frac{1}{4y_j}\right)}$$

$$= \exp\left[\frac{Q_j}{4}(w_j - y_j)\right]: \quad j = 1, 2, \ldots, p. \quad (51)$$

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” $y_j^2 = 2/w_j$ is a member of $p(y_j^2|ELSE)$ with probability $R$ as well; otherwise stay with the current $y_j^2$.

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 $g_1 = X\beta_1$ and $g_2 = X\beta_2$ be the genomic values of the $N$ individuals for each of the traits. A model could be

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \mu + g_1 + g_2 + e_1 e_2.$$  \quad (52)

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

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} | G, G_0 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, G_0 \otimes G_0\right). \quad (53)$$

where $G$ is an $N \times N$ marker-based matrix of “genomic relationships,” and

$$G_0 = \begin{bmatrix} \sigma_{g_1}^2 & \sigma_{g_{12}} \\ \sigma_{g_{12}} & \sigma_{g_2}^2 \end{bmatrix} \quad (54)$$

is a matrix containing the trait-specific genomic variances and their covariances. Specifically, from the definition of $g_1$ and $g_2$, and, assuming that $\beta_t | \sigma_{\beta_t}^2 \sim N(0, \sigma_{\beta_t}^2)$ ($t = 1, 2$),

$$\text{Var}(g_t | X, \sigma_{\beta_t}^2) = \frac{1}{p} XX' \left(\sigma_{\beta_t}^2 \sigma_{\beta_{12}} \right) = G_0 \sigma_{g_t}^2; \quad t = 1, 2, \quad (55)$$

for $G = XX' / p$ and $\sigma_{g_t}^2 = p \sigma_{\beta_t}^2$. Similarly, $\text{Cov}(g_1, g_2 | X, \sigma_{\beta_{12}}) = G_0 \sigma_{g_{12}}$, where $\sigma_{g_{12}} = p \sigma_{\beta_{12}}$ and $\sigma_{\beta_{12}}$ is the covariance between marker effects on traits 1 and 2. Let $B = \{\sigma_{\beta_t}\}$ be the $2 \times 2$ variance–covariance matrix of marker effects.
For a bivariate Bayesian LASSO model, conditionally on the \( p \times 1 \) vector \( \mathbf{v}^2 \), one has

\[
\begin{bmatrix}
\beta_1^* \\
\beta_2^*
\end{bmatrix} | \Sigma, \mathbf{v}^2 \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \otimes \mathbf{D} \right).
\]

(Hence

\[
\begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} | \Sigma, \mathbf{v}^2
\sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \otimes \mathbf{X} \mathbf{D} \mathbf{X}' \right)
= N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \otimes \mathbf{X} \mathbf{D} \mathbf{X}' \right).
\]

Let \( \mathbf{C}_{\text{Cond}} = \Sigma \otimes \mathbf{X} \mathbf{D} \mathbf{X}' \). Further,

\[
\begin{align*}
E \left( \begin{bmatrix}
\mathbf{y}_1 \\
\mathbf{y}_2
\end{bmatrix} | \Sigma, \mathbf{v}^2, \mathbf{R}_0 \right) &= \begin{bmatrix} 1 \mu_1 \\ 1 \mu_2 \end{bmatrix},
\end{align*}
\]

and

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

After assigning a flat prior to each of \( \mu_1 \) and \( \mu_2 \), standard results give that posterior distribution of the genotypic values given \( \Sigma, \mathbf{v}^2, \mathbf{R}_0 \) is normal, with mean vector

\[
\begin{align*}
E \left( \begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} | \Sigma, \mathbf{v}^2, \mathbf{R}_0, \mathbf{y} \right) &= \mathbf{C}_{\text{Cond}} \mathbf{V}_{\text{Cond}}^{-1} \left( \begin{bmatrix} \mathbf{y}_1 - 1 \mu_1 \\ \mathbf{y}_2 - 1 \mu_2 \end{bmatrix} \right) = \begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \end{bmatrix},
\end{align*}
\]

where

\[
\begin{align*}
\begin{bmatrix}
\hat{\mu}_1 \\
\hat{\mu}_2
\end{bmatrix} &= \left( \begin{bmatrix} 1' & 0 \\ 0 & 1' \end{bmatrix} \mathbf{V}_{\text{Cond}}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \\
&\times \left( \begin{bmatrix} 1' & 0 \\ 0 & 1' \end{bmatrix} \mathbf{V}_{\text{Cond}}^{-1} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right)^{-1}.
\end{align*}
\]

Further (Henderson 1975)

\[
\begin{align*}
\text{Var} \left( \begin{bmatrix}
g_1 \\
g_2
\end{bmatrix} | \Sigma, \mathbf{v}^2, \mathbf{R}_0, \mathbf{y} \right) &= \mathbf{C}_{\text{Cond}} - \mathbf{C}_{\text{Cond}} \mathbf{V}_{\text{Cond}}^{-1} \mathbf{C}_{\text{Cond}} + \mathbf{C}_{\text{Cond}} \mathbf{V}_{\text{Cond}}^{-1} (1' \mathbf{V}_{\text{Cond}}^{-1} 1) 1' \mathbf{V}_{\text{Cond}}^{-1} \mathbf{C}_{\text{Cond}} \\
&= \mathbf{C}_{\text{Cond}} - \mathbf{C}_{\text{Cond}} [\mathbf{V}_{\text{Cond}}^{-1} - \mathbf{V}_{\text{Cond}}^{-1} (1' \mathbf{V}_{\text{Cond}}^{-1} 1) 1' \mathbf{V}_{\text{Cond}}^{-1}] \mathbf{C}_{\text{Cond}}.
\end{align*}
\]

Hence, draws from the conditional posterior distribution of \( \mathbf{g} = [g_1 \ g_2]' \) given \( \Sigma, \mathbf{v}^2 \), 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 $\Sigma, v^2$ and $R_0$, the vector $[\beta_1^*, \beta_2^*, g_1^*, g_2^*]^T$ has a multivariate normal distribution, and let $\beta^* = [\beta_1^*, \beta_2^*, g_1^*, g_2^*]^T$. Hence

$$E(\beta^*|\Sigma, v^2, R_0, y) = E_{g|\Sigma, v^2, R_0, y}[\beta^*|g, \Sigma, v^2, R_0]$$

$$= E_{g|\Sigma, v^2, R_0, y}[\text{Cov}(\beta^*, g') (\Sigma \otimes X'DX')^{-1} g']$$

$$= \text{Cov}(\beta^*, g') (\Sigma \otimes XDX')^{-1} \bar{g} = \bar{\beta}^*.$$  \hspace{1cm} (62)

Now,

$$\text{Cov}(\beta^*, g') = \text{Cov} \left( \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ g_1^* \\ g_2^* \end{bmatrix} \right) = B \otimes X',$$  \hspace{1cm} (63)

so

$$\bar{\beta}^* = (B \otimes X') (\Sigma \otimes XDX')^{-1} = (B \Sigma^{-1} \otimes X'DX') \bar{g}$$  \hspace{1cm} (64)

Similarly

$$\text{Var}(\beta^*|\Sigma, v^2, R_0, y) = \text{Var}_{g|\Sigma, v^2, R_0, y}[E(\beta^*|g, \Sigma, v^2, R_0)]$$

$$+ E_{g|\Sigma, v^2, R_0, y} \text{Var}[\beta^*|g, \Sigma, v^2, R_0]$$

$$= \text{Var}_{g|\Sigma, v^2, R_0, y}[(B \Sigma^{-1} \otimes X'DX') \bar{g}]$$

$$+ (B \otimes I) (\Sigma \otimes XDX')^{-1} (B \otimes I)$$

$$= B \Sigma^{-1} B \otimes (XDX')^{-1}.$$  \hspace{1cm} (65)

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

Despite 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 reweighted set of linear “mixed model equations,” for which extremely efficient solvers exist. It is assumed that “good” estimates of $R_0$ (the residual covariance matrix) and of $B$ (the $T \times T$ variance-covariance matrix of markers effects) are available. From (8) $\Sigma = B/[4(T + 1)]$, e.g., for $T = 3$ then $\Sigma = B/16$; hence, an assessment of the scale matrix of the MLAP distribution is easily available.

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

$$\log[p(\theta|R_0, \Sigma, \text{DATA})]$$

$$= -\frac{1}{2} \left( y' - (I_3 \otimes I_N) \mu - (I_3 \otimes X) \beta' \right)' [R_0^{-1} \otimes I_N] \left( y' - (I_3 \otimes I_N) \mu - (I_3 \otimes X) \beta' \right)$$

328  D. Gianola and R. L. Fernando
\[-\frac{1}{2} \sum_{j=1}^{p} \sqrt{\beta_j'\Sigma^{-1}\beta_j} = L(\theta) \]  

(66)

Let

\[ L(\theta) = L_{lik}(\mu, \beta^*) + L_{prior}(\beta), \]  

(67)

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

\[ \frac{\partial L_{lik}(\mu, \beta^*)}{\partial \mu} = (R_0^{-1} \otimes I_N') \left[ y^* - (I_3 \otimes I_N)\mu - (I_3 \otimes X)\beta^* \right]. \]  

(68)

and

\[ \frac{\partial L_{lik}(\mu, \beta^*)}{\partial \beta^*} = (R_0^{-1} \otimes X') \left[ y^* - (I_3 \otimes I_N)\mu - (I_3 \otimes X)\beta^* \right] \]  

(69)

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

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

(70)

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

\[
\frac{\partial}{\partial \beta} L_{prior}(\beta) = -\Sigma^{-1} \begin{bmatrix}
\frac{1}{m_1} (\Sigma^{11}\beta_{j1} + \Sigma^{12}\beta_{j2} + \Sigma^{13}\beta_{j3}) \\
\frac{1}{m_2} (\Sigma^{21}\beta_{j1} + \Sigma^{22}\beta_{j2} + \Sigma^{23}\beta_{j3}) \\
\frac{1}{m_3} (\Sigma^{31}\beta_{j1} + \Sigma^{32}\beta_{j2} + \Sigma^{33}\beta_{j3})
\end{bmatrix}; \ j = 1, 2, ..., p.
\]  

(71)

where \( m_j = 2\sqrt{\beta_j'\Sigma^{-1}\beta_j} \) is proportional to the Mahalanobis distance of \( \beta_j \) 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

\[
\frac{\partial}{\partial \beta} L_{prior}(\beta) = -\begin{bmatrix}
\Sigma^{-1}m_1\beta_1 \\
\Sigma^{-1}m_2\beta_2 \\
\Sigma^{-1}m_3\beta_3 \\
\cdot \\
\cdot \\
\cdot \\
\Sigma^{-1}m_p\beta_p
\end{bmatrix} = -(M^{-1} \otimes \Sigma^{-1}) \beta,
\]  

(72)

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

\[
\frac{\partial}{\partial \beta} L_{prior}(\beta^*) = -(\Sigma^{-1} \otimes M^{-1}) \beta^*.
\]  

(73)
Collecting (69) and (73),

\[
\frac{\partial}{\partial \beta^*} L(\theta) = \frac{\partial}{\partial \beta^*} L_\text{lik}(\theta) + \frac{\partial}{\partial \beta^*} L_\text{prior}(\beta^*)
\]

\[
= (R_0^{-1} \otimes X') \left[ y^* - (I_3 \otimes 1_N)\mu - (I_3 \otimes X)\beta^* \right] - (\Sigma^{-1} \otimes M^{-1})\beta^*, \tag{74}
\]

Setting (68) and (74) simultaneously to 0 produces the system of equations (not explicit)

\[
\begin{bmatrix}
R_0^{-1} \otimes N \\
R_0^{-1} \otimes X'1_N \\
(I_3 \otimes X' + (\Sigma^{-1} \otimes M^{-1}))
\end{bmatrix}
\begin{bmatrix}
\mu \\
\beta
\end{bmatrix}
= \begin{bmatrix}
(R_0^{-1} \otimes 1_N) y^* \\
(R_0^{-1} \otimes X') y^*
\end{bmatrix}. \tag{75}
\]

Expanding the equations above for \( T = 3 \) yields

\[
\begin{bmatrix}
N^2 X 1_N \\
X_1 X_1 N \\
X_2 X_1 N \\
X_2 X_1 N \\
X_1 X_1 N \\
X'_1 X_1 N
\end{bmatrix}
\begin{bmatrix}
\mu \\
\beta
\end{bmatrix}^{[b+1]}
= \begin{bmatrix}
1_N (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
1_N (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
1_N (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*) \\
X' (r_{11} y_1^* + r_{12} y_2^* + r_{13} y_3^*) \\
X' (r_{21} y_1^* + r_{22} y_2^* + r_{23} y_3^*) \\
X' (r_{31} y_1^* + r_{32} y_2^* + r_{33} y_3^*)
\end{bmatrix}^{[b]}
\times
\begin{bmatrix}
N^2 X 1_N \\
X_1 X_1 N \\
X_2 X_1 N \\
X_2 X_1 N \\
X_1 X_1 N \\
X'_1 X_1 N
\end{bmatrix}
\begin{bmatrix}
\mu \\
\beta
\end{bmatrix}^{[b]}
\begin{bmatrix}
N^2 X 1_N \\
X_1 X_1 N \\
X_2 X_1 N \\
X_2 X_1 N \\
X_1 X_1 N \\
X'_1 X_1 N
\end{bmatrix}
\begin{bmatrix}
\mu \\
\beta
\end{bmatrix}, \tag{76}
\]

where \( b \) is iterate round. Matrix \( M = \text{Diag}(m_j) \) 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 \( m_j \) values and, consequently, \( M^{-1} \) will have large diagonal elements. Hence, vectors of markers with weak effects are more strongly shrunk toward 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

\[
\text{Var} \left( \begin{bmatrix}
\mu_1 \\
\mu_3 \\
\beta_1 \\
\beta_3
\end{bmatrix} \mid R_0, \Sigma, \text{DATA} \right) \approx K_i^{-1}, \tag{77}
\]

with \( \approx \) 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

\[
y_i^{\text{complete}} = (y_i^{\text{miss}}, y_i^{\text{obs}}), \tag{78}
\]
where miss denotes a missing record, e.g., if \( T = 2 \), a record could be missing for trait 1 or for trait 2; \( y_i^{\text{obs}} \) represents the phenotypes for the traits observed in individual \( i \). The posterior predictive distribution of \( y_i^{\text{miss}} \) has density

\[
p(y_i^{\text{miss}} | y, R_0, \Sigma) = \int p(y_i^{\text{miss}} | \mu, \beta, R_0, y_i^{\text{obs}}) p(\mu, \beta, R_0, \Sigma | y) d\mu d\beta dR_0 d\Sigma,
\]

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

\[
y_i^{\text{miss}} | y, R_0, \Sigma \sim N\left( E\left( y_i^{\text{miss}} | \mu, \beta, R_0, y_i^{\text{obs}} \right), \text{Var}\left( y_i^{\text{miss}} | \mu, \beta, R_0, y_i^{\text{obs}} \right) \right)
\]

Since the sampling model is normal, for \( T = 3 \), one has

\[
E\left( y_i^{\text{miss}} | \mu, \beta, R_0, y_i^{\text{obs}} \right) = \begin{bmatrix} \delta_1 \mu_1 \\ \delta_2 \mu_2 \\ \delta_3 \mu_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \delta_4 x_i^1 \\ 0 & \delta_2 x_i^2 & 0 \\ 0 & 0 & \delta_3 x_i^3 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{bmatrix}
\]

\[+ R_0^{\text{miss,obs}} (R_0^{\text{obs,obs}})^{-1} e^{\text{obs}}, \]

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; \( R_0^{\text{obs,obs}} \) is the part of \( R_0 \) pertaining to observed phenotypes for case \( i \), and \( R_0^{\text{miss,obs}} \) is the submatrix of residual covariances between missing and observed traits. Further,

\[
\text{Var}\left( y_i^{\text{miss}} | \mu, \beta, R_0, y_i^{\text{obs}} \right) = R_0^{\text{miss,miss}} - R_0^{\text{miss,miss}} (R_0^{\text{obs,obs}})^{-1} R_0^{\text{obs,miss}}
\]

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

\[
E\left( y_{250}^{\text{miss}} | \mu, \beta, R_0, y_{250}^{\text{obs}} \right) = \mu_1 + x_{250} \beta_1^* + \begin{bmatrix} r_{12} & r_{13} \end{bmatrix} \begin{bmatrix} r_{22} & r_{23} \\ r_{32} & r_{33} \end{bmatrix}^{-1} \begin{bmatrix} e_{2,250} \\ e_{3,250} \end{bmatrix},
\]

and

\[
\text{Var}\left( y_{250}^{\text{miss}} | \mu, \beta, R_0, y_{250}^{\text{obs}} \right) = r_{11} - \begin{bmatrix} r_{12} & r_{13} \end{bmatrix} \begin{bmatrix} r_{22} & r_{23} \\ r_{32} & r_{33} \end{bmatrix}^{-1} \begin{bmatrix} r_{12} \\ r_{13} \end{bmatrix}.
\]

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).