OXFORD GENETICS

GENETICS, 2024, iyae179
https://doi.org/10.1093/genetics/iyae179
Genomic Data Analyses in Biobanks

Megavariate methods capture complex genotype-by-environment interactions

Alencar Xavier [1,2,*] Daniel Runcie [3] David Habier[1]

[1]Corteva Agrisciences, Seed Product Development, 8305 NW 62nd Ave, Johnston, IA 50131, USA
[2]Purdue University, Department of Agronomy, 915 Mitch Daniels Blvd, West Lafayette, IN 47907, USA
[3]University of California Davis, Department of Plant Sciences, One Shield Ave, Davis, CA 95616, USA

*Corresponding author: Corteva Agrisciences, Seed Product Development, 8305 NW 62nd Ave, Johnston, IA 50131, USA. Email: alencar.xavier@corteva.com

Genomic prediction models that capture genotype-by-environment (GxE) interaction are useful for predicting site-specific performance by leveraging information among related individuals and correlated environments, but implementing such models is computationally challenging. This study describes the algorithm of these scalable approaches, including 2 models with latent representations of GxE interactions, namely MegaLMM and MegaSEM, and an efficient multivariate mixed-model solver, namely Pseudo-expectation Gauss–Seidel (PEGS), fitting different covariance structures [unstructured, extended factor analytic (XFA), Heteroskedastic compound symmetry (HCS)]. Accuracy and runtime are benchmarked on simulated scenarios with varying numbers of genotypes and environments. MegaLMM and PEGS-based XFA and HCS models provided the highest accuracy under sparse testing with 100 testing environments. PEGS-based unstructured model was orders of magnitude faster than restricted maximum likelihood (REML) based multivariate genomic best linear unbiased predictions (GBLUP) while providing the same accuracy. MegaSEM provided the lowest runtime, fitting a model with 200 traits and 20,000 individuals in ~5 min, and a model with 2,000 traits and 2,000 individuals in less than 3 min. With the genomes-to-fields data, the most accurate predictions were attained with the univariate model fitted across environments and by averaging environment-level genomic estimated breeding values (GEBVs) from models with HCS and XFA covariance structures.

Keywords: accuracy; genomic prediction; multivariate models; matrix decomposition

Introduction

Multienvironment trials (METs) are the main source of data for plant breeding decision-making. Genotype-by-environment (GxE) interactions occur when the performance of a genotype varies across trials due to different environmental conditions. Predicting genotype performance within environments is essential for enhanced adaptation and selection of superior genotypes. Understanding how genotypes perform in diverse environments allows breeders to tailor cultivars to specific conditions or select genotypes that perform consistently well across different environments. This leads to more effective development of cultivars and ensures new cultivars thrive in the intended conditions (Elias *et al.* 2016).

Genomic prediction (Meuwissen *et al.* 2001) is used to predict genotype performance within environments. It exploits genomic relationships between genotypes to predict breeding values (Habier *et al.* 2007), which accelerates genetic progress by increasing the accuracy of predictions and shortening breeding cycles, while it also allows breeders to reduce phenotyping costs (Crossa *et al.* 2021). Genomic prediction models that fit genetic correlations between pairs of environments, thereby capturing GxE interaction, can be more accurate than univariate (UV) models that fit phenotypes of each environment independently (Xavier and Habier 2022).

Genomic prediction models that account for GxE interactions are computationally demanding (Heslot *et al.* 2014). They are more complex than UV genomic models due to many covariance parameters that need to be estimated for solving large and dense mixed-model equations (Hardner 2017; Martini *et al.* 2020). The most complex one is the unstructured model that estimates a different covariance for each pair of environments (Bustos-Korts *et al.* 2016; Crossa *et al.* 2022). The computational burden of such a model is further exacerbated by the progress in high-throughput phenotyping, which increases the number of agronomic traits and environments substantially. One solution for reducing model complexity is fitting compound symmetry models that assume a constant genetic correlation for all pairs of environments (Cuevas *et al.* 2016). Another solution is fitting GxE interaction kernels that compute the Hadamard product of genetic and environment covariance matrices and scale it by a single variance (Jarquín *et al.* 2014). Often, the analysis of MET even follows simpler parameterizations by using Finlay–Wilkinson approaches or GGE models (Malosetti *et al.* 2013). All of these solutions may have lower accuracy in capturing GxE interactions and predictions than unstructured models. Thus, computationally efficient

Received on 18 September 2024; accepted on 26 October 2024

methods for estimating parameters and solving mixed-model equations of these complex models are needed when the number of environments and traits is large.

The purpose of this study is to review computationally efficient methods that can be utilized for genomic prediction within environments while capturing complex GxE patterns. We rate the scalability of these methods for applying them to varying numbers of genotypes, markers, and environments. Benchmarks of accuracy and runtime are performed on multiple simulated scenarios and real data.

Materials and Methods

In environment-specific models, phenotypes in environment *k* are modeled as

$$\mathbf{y_k} = \mathbf{1}\boldsymbol{\mu}_\mathbf{k} + \mathbf{g_k} + \mathbf{e_k} \tag{1}$$

with variance

$$\begin{aligned} Var(\mathbf{y_k}) &= \mathbf{V_k} \\ &= \mathbf{G_k} + \mathbf{R_k}, \\ Var(\mathbf{g_k}) &= \mathbf{G_k}, \\ Var(\mathbf{e_k}) &= \mathbf{R_k}, \end{aligned} \tag{2}$$

where the vector of phenotypes ($\mathbf{y_k}$) is modeled by the overall mean ($\boldsymbol{\mu}_\mathbf{k}$), a vector of genetic signal ($\mathbf{g_k}$) and a vector of residuals ($\mathbf{e_k}$). The phenotypic covariance matrix $\mathbf{V_k}$ is the sum of genetic ($\mathbf{G_k}$) and residual ($\mathbf{R_k}$) covariance matrices. In an additive model, the genetic covariance matrix is calculated as

$$\mathbf{G_k} = \mathbf{Z_k}\mathbf{Z_k}'\boldsymbol{\sigma}^2_{\beta(\mathbf{k})}, \tag{3}$$

where $\boldsymbol{\sigma}^2_{\beta(\mathbf{k})}$ denotes the variance of marker effects and $\mathbf{Z}$ is a matrix that contains the molecular allele dosage coded as 0, 1, or 2 for different genotypes in rows and different marker loci in columns of the matrix. Residuals in each environment *k* are assumed independent and identically distributed as $\mathbf{R_k} = \mathbf{I}\boldsymbol{\sigma}^2_{\mathbf{e(k)}}$. The genetic term can be described as a linear combination of marker effects (Habier *et al.* 2007) as

$$\mathbf{g_k} = \mathbf{Z_k}\boldsymbol{\beta}_\mathbf{k}, \tag{4}$$

where $Var(\boldsymbol{\beta}_\mathbf{k}) = \mathbf{I}\boldsymbol{\sigma}^2_{\beta(\mathbf{k})}$. In MET analyses, the covariance between a pair of environments is denoted as $Cov(\boldsymbol{\beta}_k, \boldsymbol{\beta}_{k'})$, which is a function of the GxE correlation between environments *k* and *k*′, $\boldsymbol{\rho}_{\beta(\mathbf{k},\mathbf{k}')}$, as

$$\boldsymbol{\sigma}_{\beta(\mathbf{k},\mathbf{k}')} = \boldsymbol{\rho}_{\beta(\mathbf{k},\mathbf{k}')}\boldsymbol{\sigma}_{\beta(\mathbf{k})}\boldsymbol{\sigma}_{\beta(\mathbf{k}')}. \tag{5}$$

Realistically, every environment has its unique signal ($\boldsymbol{\beta}_\mathbf{k}$) and every pair of environments has a unique correlation (Falconer 1952). Approaches designed to accommodate such assumptions are presented next. Collectively, these correspond to efficient (1) models, (2) parameterizations, and (3) solvers.

(1) Efficient models

Univariate by environment

By treating environments as independent, the UV model is the simplest approach to obtaining environment-specific predictions. They are obtained by fitting location-year combinations from either raw or spatially adjusted phenotypes (Möhring and Piepho 2009; Piepho *et al.* 2012). Accurate predictions can be achieved when genotypes in the model are related, but accuracy can be constrained by the dataset size (Xu 2003) and genetic scope (Habier *et al.* 2013).

Multivariate across environment

Environments are treated as different traits that are fitted with an unstructured genetic covariance matrix (Falconer 1952; Falconer and Mackay 1983). This requires the estimation of genetic variances and covariances between pairs of environments [equation (5)].

Simplified canonical transformation

For complete data across environments, i.e. balanced without missing phenotypes, canonical transformation (CT) is an efficient method to fit all environments simultaneously, while capturing the shared genetic information across correlated environments (Meyer 1985; Konstantinov and Erasmus 1993). The traditional CT method (see Appendix A) iterates through transforming trait phenotypes and solving transformed variance components until convergence.

A simplified canonical transformation (SCT) that avoids the iterative estimation of covariances is described below. Phenotypes are transformed once by singular value decomposition (SVD)

$$\begin{aligned} \mathbf{Y} &= \mathbf{UDV}' \\ &= \mathbf{FV}' \end{aligned} \tag{6}$$

where $\mathbf{Y} = \{\mathbf{y_1}, \ldots, \mathbf{y_K}\}$ is a matrix of phenotypes with rows and columns representing observations and traits, respectively, *K* is the number of environments, $\mathbf{F} = \mathbf{UD}$ is a latent traits matrix that equals the principal components of $\mathbf{Y}$, and $\mathbf{V}$ is a rotation matrix. The latent traits are orthogonal with $cov(\mathbf{f_i}, \mathbf{f_j}) = 0$ for every pair of latent traits **i** and **j**. Subsequently, each latent trait is fitted using a UV model

$$\mathbf{f_k} = \mathbf{1}\boldsymbol{\mu}_\mathbf{k} + \mathbf{Z_k}\boldsymbol{\gamma}_\mathbf{k} + \mathbf{e_k}, \tag{7}$$

and estimated breeding values of the original traits are calculated by

$$\begin{aligned} \hat{\mathbf{G}} &= \mathbf{Z}\hat{\boldsymbol{\Gamma}}\mathbf{V}' \\ &= \mathbf{Z}\hat{\mathbf{B}}, \end{aligned} \tag{8}$$

where $\hat{\mathbf{G}} = \{\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_K\}$, $\hat{\mathbf{B}} = \hat{\boldsymbol{\Gamma}}\mathbf{V}$, $\hat{\boldsymbol{\Gamma}} = \{\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_K\}$, and $\mathbf{V}$ is rotating marker effects back to the natural scale of the phenotypes.

MegaLMM

MegaLMM (Runcie *et al.* 2021) extends the decomposition of SCT with a more parsimonious latent representation of the phenotypes, with the addition of trait-specific model terms (see Appendix B). This enables efficient handling of missing values in $\mathbf{Y}$ while permitting the model to accommodate more traits. Latent spaces are inferred from a stochastic matrix decomposition of $\mathbf{Y}$ based on the following statistical model:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\Lambda} + \mathbf{J}. \tag{9}$$

The matrix of phenotypes is decomposed into latent spaces $\mathbf{F}$ rotated by $\boldsymbol{\Lambda}$ and residuals $\mathbf{J}$. This residual matrix $\mathbf{J}$ contains genetic signal ($\mathbf{Z}\boldsymbol{\Delta}$) not captured by $\mathbf{F}\boldsymbol{\Lambda}$ and true error ($\mathbf{E}$), thus $\mathbf{J} = \mathbf{Z}\boldsymbol{\Delta} + \mathbf{E}$. In this approach, each trait is fitted by a UV model as

$$\mathbf{y_k} = \mathbf{1}\mu_\mathbf{k} + \mathbf{F}\boldsymbol{\lambda}_\mathbf{k} + \mathbf{Z_k}\boldsymbol{\delta}_\mathbf{k} + \mathbf{e_k}, \tag{10}$$

and, then, each latent space is modeled by

$$\mathbf{f_l} = \mathbf{1}\mu_\mathbf{l} + \mathbf{Z}\boldsymbol{\gamma}_\mathbf{l} + \mathbf{e_l}. \tag{11}$$

The shared genetic signal is captured by $\mathbf{F}\boldsymbol{\lambda}_\mathbf{k}$, while the environment-specific genetic signal is captured by $\mathbf{Z_k}\boldsymbol{\delta}_\mathbf{k}$. Multivariate marker effects are estimated by $\hat{\boldsymbol{\beta}}_\mathbf{k} = \boldsymbol{\Gamma}\hat{\boldsymbol{\lambda}}_\mathbf{k} + \hat{\boldsymbol{\delta}}_\mathbf{k}$, so that the complete matrix of estimated breeding values can be calculated with

$$\begin{aligned} \hat{\mathbf{G}} &= \mathbf{Z}\boldsymbol{\Gamma}\hat{\boldsymbol{\Lambda}} + \mathbf{Z}\hat{\boldsymbol{\Delta}} \\ &= \mathbf{Z}(\boldsymbol{\Gamma}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\Delta}}) \\ &= \mathbf{Z}\hat{\mathbf{B}}. \end{aligned} \tag{12}$$

The latent spaces are strongly shrunken according to their relative importance (Runcie *et al.* 2021). The original implementation of MegaLMM uses Markov Chain Monte Carlo (MCMC) for estimating $\hat{\mathbf{F}}$ and $\hat{\boldsymbol{\Lambda}}$ by alternating between the conditional models $(\mathbf{Y} \mid \mathbf{F}) = \mathbf{F}\boldsymbol{\Lambda} + \mathbf{J}$ and $(\mathbf{Y}' \mid \boldsymbol{\Lambda}) = \boldsymbol{\Lambda}\mathbf{F} + \mathbf{J}'$ for updating the coefficients.

Structural equation models

Structural equation models (SEMs) are used for modeling directionally correlated response variables (Gianola and Sorensen 2004) and causal trait networks (Valente *et al.* 2013). For MET analyses, a fully connected SEM fits the phenotypes of environment k as a function of phenotypes at other environments and model terms specific to environment k:

$$\mathbf{y_k} = \mathbf{1}\mu_\mathbf{k} + \mathbf{Y}\boldsymbol{\psi}_\mathbf{k} + \mathbf{Z_k}\boldsymbol{\delta}_\mathbf{k} + \mathbf{e_k}, \tag{13}$$

where $\mu_\mathbf{k}$ is the intercept, $\boldsymbol{\psi}_\mathbf{k}$ is a vector of length k that quantifies the linear associations between phenotypes of trait k and phenotypes from other traits ($\psi_\mathbf{k} = 0$), and $\boldsymbol{\delta}_\mathbf{k}$ is a vector of environment-specific marker effects. Marker effects are estimated from

$$\hat{\mathbf{B}} = (\mathbf{I} - \hat{\boldsymbol{\Psi}})^{-1}\hat{\boldsymbol{\Lambda}}, \tag{14}$$

where $\hat{\boldsymbol{\Psi}} = \{\hat{\boldsymbol{\psi}}_\mathbf{1}, \hat{\boldsymbol{\psi}}_\mathbf{2}, \ldots, \hat{\boldsymbol{\psi}}_\mathbf{K}\}$ and $\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\delta}}_\mathbf{1}, \hat{\boldsymbol{\delta}}_\mathbf{2}, \ldots, \hat{\boldsymbol{\delta}}_\mathbf{K}\}$.

The SEM solution is straightforward for balanced datasets with a small number of traits. SEM differs from the MegaLMM model by explicitly parameterizing phenotypic traits ($\mathbf{Y}\boldsymbol{\Psi}$) in the model as opposed to using a latent representation ($\mathbf{F}\boldsymbol{\lambda}$).

MegaSEM

An intermediate parameterization between SEM and MegaLMM, here referred to as MegaSEM, can be achieved by utilizing a matrix of UV genomic estimated breeding values (GEBVs) from all traits to parameterize the model for every trait. The matrix of UV GEBVs is given by $\mathbf{G_0} = \mathbf{ZB_0}$, using UV marker effects $\mathbf{B_0} = \{\mathbf{b_{0(1)}}, \mathbf{b_{0(2)}}, \ldots, \mathbf{b_{0(K)}}\}$. The matrix of UV GEBVs is then decomposed by SVD as

$$\begin{aligned} \mathbf{G_0} &= \mathbf{U_0 D_0 V_0'} \\ &= \mathbf{F_0 V_0'}. \end{aligned} \tag{15}$$

Once principal components ($\mathbf{F_0} = \mathbf{U_0 D_0}$) have been computed, MegaSEM fits each trait as

$$\mathbf{y_k} = \mathbf{1}\mu_\mathbf{k} + \mathbf{F_0}\boldsymbol{\alpha}_\mathbf{k} + \mathbf{Z_k}\boldsymbol{\delta}_\mathbf{k} + \mathbf{e_k}, \tag{16}$$

which has the same form as equation (10). Similar to the MegaLMM model, the shared genetic signal is captured by $\mathbf{F_0}\boldsymbol{\alpha}_\mathbf{k}$ and environment-specific genetic signal is captured by $\mathbf{Z_k}\boldsymbol{\delta}_\mathbf{k}$. Estimated breeding values are calculated by

$$\begin{aligned} \hat{\mathbf{G}} &= \mathbf{F_0}\hat{\mathbf{A}} + \mathbf{Z}\hat{\boldsymbol{\Delta}} \\ &= \mathbf{ZB_0V_0}\hat{\mathbf{A}} + \mathbf{Z}\hat{\boldsymbol{\Delta}} \\ &= \mathbf{Z}(\mathbf{B_0V_0}\hat{\mathbf{A}} + \hat{\boldsymbol{\Delta}}) \\ &= \mathbf{Z}\hat{\mathbf{B}}. \end{aligned} \tag{17}$$

Note that $\mathbf{F_0} = \mathbf{ZB_0V_0}$, because

$$\begin{aligned} \mathbf{G_0} &= \mathbf{ZB_0} \\ \mathbf{F_0V_0'} &= \mathbf{ZB_0} \\ \mathbf{F_0V_0'V_0} &= \mathbf{ZB_0V_0} \\ \mathbf{F_0} &= \mathbf{ZB_0V_0}, \end{aligned} \tag{18}$$

as $\mathbf{V_0'V_0} = \mathbf{I}$. As shown in equation (18), the estimated marker effects for trait k, $\hat{\boldsymbol{\beta}}_\mathbf{k}$, are a linear combination of marker effects from all environments estimated by UV analyses in $\mathbf{B_0}$, and environment-specific effects $\hat{\boldsymbol{\alpha}}_k$ and $\hat{\boldsymbol{\delta}}_k$, i.e. $\hat{\boldsymbol{\beta}}_\mathbf{k} = \mathbf{B_0V_0'}\hat{\boldsymbol{\alpha}}_\mathbf{k} + \hat{\boldsymbol{\delta}}_\mathbf{k}$.When all principal components of $\mathbf{G_0}$ are utilized (i.e. no dimensionality reduction), the MegaSEM model can be simplified further by

$$\mathbf{y_k} = \mathbf{1}\mu_\mathbf{k} + \mathbf{F_0}\boldsymbol{\alpha}_\mathbf{k} + \mathbf{e_k}, \tag{19}$$

where the environment-specific term can be omitted if the model is parameterized with all linear combinations of environments, which yields $\hat{\mathbf{B}} = \mathbf{B_0V_0}\hat{\mathbf{A}}$.

The computational cost of MegaSEM consists of running a UV model twice, first to estimate $\mathbf{B_0}$ and then to calculate $\hat{\mathbf{B}}$, in addition to the SVD of $\mathbf{G_0}$.

(2) Efficient parameterizations

Kernels and rotations

Kernel-based genomic prediction is used to parameterize linear and nonlinear relationships (de Los Campos *et al.* 2010, 2013; Montesinos-López *et al.* 2021). Rotation of kernels by spectral, eigenvalue decomposition (EVD), or singular-value decomposition (SVD), enables solving such models by a Gauss–Seidel (GS) algorithm (Legarra and Misztal 2008). Rotations are also useful when the number of parameters far exceeds the number of genotypes because they can reduce the dimensionality of the problem substantially (Ødegård *et al.* 2018; Xavier and Habier 2022). Genomic prediction based on a kernel $\mathbf{K}$ can be described as

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{g} + \mathbf{e}, \\ \mathbf{g} &\sim N(0, \mathbf{K}\sigma_g^2), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma_e^2). \end{aligned} \tag{20}$$

The kernel, which describes genetic relationships, can be decomposed by EVD (Thompson and Shaw 1990) as

$$\begin{aligned} \mathbf{K} &= \mathbf{UD^2U'} \\ &= (\mathbf{UD})(\mathbf{UD})' \\ &= \mathbf{QQ'}, \end{aligned} \tag{21}$$

where $\mathbf{Q} = \mathbf{UD}$. The model can be reparameterized as

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{Q}\boldsymbol{\alpha} + \mathbf{e}, \\ \boldsymbol{\alpha} &\sim N(0, \mathbf{I}\sigma_g^2), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma_e^2), \end{aligned} \tag{22}$$

where genetic effects are fitted as $\mathbf{g} = \mathbf{Q}\boldsymbol{\alpha}$ with variance–covariance matrix $Var(\mathbf{g}) = \mathbf{QQ}'\sigma_g^2 = \mathbf{K}\sigma_g^2$ and σ_g^2 is the genetic variance. To calculate predictions for individuals who were not in the training dataset (TS) used in the analysis, a rotation matrix is defined as

$$\mathbf{R} = \mathbf{UD}^{-1}. \tag{23}$$

Let $\mathbf{K_{PS,TS}}$ be the kernel that contains the genetic relationships between prediction (PS) and training dataset (TS) individuals. Then, the matrix $\mathbf{Q_{PS,TS}}$ can be calculated as

$$\mathbf{Q_{PS\,|\,TS}} = \mathbf{K_{PS,TS}R_{TS}} \tag{24}$$

and estimated breeding values can be obtained by

$$\hat{\mathbf{g}}_{\mathbf{PS}} = \mathbf{Q_{PS\,|\,TS}}\hat{\boldsymbol{\alpha}}. \tag{25}$$

For verification, it can be shown that $\mathbf{Q} = \mathbf{KR} = \mathbf{UD}$.

Rotation for reducing dimensions

When fitting a linear additive model ($\mathbf{g} = \mathbf{Z}\boldsymbol{\beta}$) with a large number of parameters ($p \gg n$), the kernel trick can be used to reduce model dimensionality by kernalizing the genomic information with $\mathbf{K} = \mathbf{ZZ}'$, and then decomposing $\mathbf{K}$ via EVD [equation (21)]. Marker effects are calculated with

$$\boldsymbol{\beta} = \mathbf{Z}'\mathbf{R}\boldsymbol{\alpha}, \tag{26}$$

where the rotation matrix [$\mathbf{R}$, equation (23)] originates from the decomposition of $\mathbf{K}$. This approach reduces the number of parameters from p to n. When n is also large, the rotation matrix $\tilde{\mathbf{R}}$ can be created from the subset ($\tilde{n} < n$) as described in equation (23). The full design matrix is created as

$$\begin{aligned} \mathbf{Q}_{\mathbf{n}\,|\,\tilde{\mathbf{n}}} &= \mathbf{K}_{\mathbf{n},\tilde{\mathbf{n}}}\tilde{\mathbf{R}} \\ &= \mathbf{Z}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}} \end{aligned} \tag{27}$$

where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{R}}$ herein represent the genotypes and rotation matrix generated from a subset of individuals. The dimensionality of the matrix is hereby reduced to less than the original number of parameters and observations. Alternatively, the sparse inversion of $\mathbf{K}$ is another computationally efficient way to solve kernel models (see Appendix D).

Rotations using SVD

The kernel rotation parameterizations aforementioned, $\mathbf{Q}$ and $\mathbf{Q}_{\mathbf{n}\,|\,\tilde{\mathbf{n}}}$ can also be derived directly by SVD of $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{UDV}', \tag{28}$$

which yields the same $\mathbf{U}$ and $\mathbf{D}$ from equation (21). Principal components are obtained either with $\mathbf{Q} = \mathbf{UD}$ or $\mathbf{Q} = \mathbf{ZV}$. Subset rotation [equation (27)] are obtained with $\mathbf{Q}_{\mathbf{n}\,|\,\tilde{\mathbf{n}}} = \mathbf{Z_n V_{\tilde{n}}}$, where $\mathbf{V_{\tilde{n}}}$ comes from the SVD of a population subset. Models parameterized by $\mathbf{Q}\boldsymbol{\alpha}$ [equation (22)] using SVD recover the marker effects [equation (4)] with $\boldsymbol{\beta} = \mathbf{V}'\boldsymbol{\alpha}$.

Diagonalization

Diagonalization refers to converting a dense matrix into a diagonal matrix, which makes it useful for genetic relationship models. Using EVD for the decomposition of $\mathbf{K}$ as in equation (21), the genetic term $\mathbf{g}$ of the model in equation (20) can be written as a regression of eigenvectors, $\mathbf{U}$, resulting in

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{U}\boldsymbol{\theta} + \mathbf{e}, \\ \boldsymbol{\theta} &\sim N(0, \mathbf{D}^2\sigma_g^2), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma_e^2). \end{aligned} \tag{29}$$

Note that the covariance matrix for $\boldsymbol{\theta}$ is $Var(\boldsymbol{\theta}) = \mathbf{D}^2\sigma_g^2$, whereas the covariance matrix for $\boldsymbol{\alpha}$ in equation (22) is $Var(\boldsymbol{\alpha}) = \mathbf{I}\sigma_g^2$. Exploiting $\mathbf{U}'\mathbf{U} = \mathbf{I}$, equation (29) can be further transformed into

$$\begin{aligned} \mathbf{U}'\mathbf{y} &= \mathbf{U}'\mathbf{1}\mu + \mathbf{U}'\mathbf{U}\boldsymbol{\theta} + \mathbf{U}'\mathbf{e} \\ \mathbf{U}'\mathbf{y} &= \mathbf{U}'\mathbf{1}\mu + \boldsymbol{\theta} + \mathbf{e}, \end{aligned} \tag{30}$$

where the residual variances are unaffected because $Var(\mathbf{e}) = \mathbf{U}'\mathbf{U}\sigma_e^2 = \mathbf{I}\sigma_e^2$. Computational advantages of equation (30) are the sparsity of the mixed-model equations and a reduction in dimensionality by not using all eigenvectors. For the multivariate (MV) case, diagonalization makes it feasible to solve variance components by REML using commercial software (Gilmour *et al.* 2017), but only when data are balanced (i.e. all genotypes are observed in all environments). This approach has become particularly useful for genome-wide association methods, both UV and MV (Zhou and Stephens 2012, 2014).

(3) Efficient solvers

Efficient univariate solver

An efficient algorithm for solving mixed-model equations is called the pseudo-expectation Gauss–Seidel (PEGS) method (Xavier and Habier 2022). PEGS is iterative based on the pseudo-expectation (PE) estimator of variance components (Schaeffer 1986) along with the estimation of coefficients using GS residual updating (Legarra and Misztal 2008). A boost in convergence is achieved by randomizing the order in which coefficients are updated within each iteration (Ma *et al.* 2015). PE is an approximation of REML (VanRaden and Jung 1988) that provides unbiased and invariant variance components (Xavier and Habier 2022). In terms of implementation, PEGS resembles a non-MCMC version of the Bayesian Gibbs sampler. It estimates breeding values from a single nuclear polymorphism (SNP)-best linear unbiased prediction (BLUP) model, or ridge regression, as

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{g} + \mathbf{e} \\ &= \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}, \end{aligned} \tag{31}$$

where $\boldsymbol{\beta} \sim N(0, \mathbf{I}\sigma_\beta^2)$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$. The variance components are estimated as

$$\hat{\sigma}_{\boldsymbol{\beta}}^2 = \frac{\tilde{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}{\mathbf{Tr}(\mathbf{Z}'\mathbf{SZ})}, \tag{32}$$

$$\hat{\sigma}_e^2 = \frac{\mathbf{y}'\hat{\mathbf{e}}}{n-1}, \tag{33}$$

where $\mathbf{S}$ is the fixed effect absorption matrix defined as $\mathbf{S} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. When the intercept is the only fixed effect,

$\tilde{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{y} - \bar{\mathbf{y}})$ and $\mathbf{Tr}(\mathbf{Z}'\mathbf{S}\mathbf{Z}) = \sum_{j=1}^{J} \sum_{i=1}^{I} (\mathbf{z_{ij}} - \bar{\mathbf{z}}_{\mathbf{j}})^{\mathbf{2}}$, where $\mathbf{z_{ij}}$ denotes the allele dosage of genotype i and $\bar{\mathbf{z}}_{\mathbf{j}}$ denotes the average allele dosage, both at locus j. The marker effects are solved with

$$\hat{\beta}_{\mathbf{j}}^{(t+1)} = \frac{\mathbf{z_j}'\mathbf{e} + \hat{\beta}_{\mathbf{j}}^{t}\mathbf{z_j}'\mathbf{z_j}}{\mathbf{z_j}'\mathbf{z_j} + \hat{\sigma}_e^2\hat{\sigma}_\beta^{-2}}, \tag{34}$$

where $\mathbf{z_j}$ is a vector of allele dosage at marker j. This is followed by the update of the residuals as

$$\hat{\mathbf{e}}^{(t+1)} = \hat{\mathbf{e}}^{(t)} - \mathbf{z_j}'(\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(t+1)} - \hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(t)}). \tag{35}$$

Similarly, the intercept updates, followed by the residual update, are achieved with

$$\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} + n^{-1}\sum_{i=1}^{I}\hat{\mathbf{e}}_i^{(t)}, \tag{36}$$

$$\hat{\mathbf{e}}^{(t+1)} = \hat{\mathbf{e}}^{(t)} - (\mu^{(t+1)} - \mu^{(t)}). \tag{37}$$

This solver is suitable for estimating coefficients and variance components for UV models and different MV parameterizations, including those that do not explicitly involve the computation of covariances, such as CT, SEMs, and MegaSEM.

Efficient multivariate solver

In unstructured MV models (Xavier and Habier 2022), the genetic covariance of a pair of environments k and k', and the residual variance for environment k, have the following solution using PE:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}(\mathbf{k},\mathbf{k}')} = \frac{\tilde{\boldsymbol{\beta}}_{\mathbf{k}}'\hat{\boldsymbol{\beta}}_{\mathbf{k}'} + \tilde{\boldsymbol{\beta}}_{\mathbf{k}'}'\hat{\boldsymbol{\beta}}_{\mathbf{k}}}{\mathbf{Tr}(\mathbf{Z}_{\mathbf{k}}'\mathbf{S}_{\mathbf{k}}\mathbf{Z}_{\mathbf{k}}) + \mathbf{Tr}(\mathbf{Z}_{\mathbf{k}'}'\mathbf{S}_{\mathbf{k}'}\mathbf{Z}_{\mathbf{k}'})}, \tag{38}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{e}(\mathbf{k})} = \frac{\mathbf{y}_{\mathbf{k}}'\hat{\mathbf{e}}_{\mathbf{k}}}{n_k - 1}, \tag{39}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}(\mathbf{k},\mathbf{k}')}$ denotes the estimated variance–covariance matrix of marker effects in different environments and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}(\mathbf{k})}$ is a matrix that contains estimated residual variances for different environments on its diagonal.

The MV solution for updating coefficients for the jth marker is

$$\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(t+1)} = (\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'\dot{\mathbf{Z}}_{\mathbf{j}} + \hat{\boldsymbol{\Sigma}}_\beta^{-1})^{-1}\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'(\dot{\mathbf{Z}}_{\mathbf{j}}\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(t)} + \hat{\mathbf{e}}), \tag{40}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{j}}$ is a vector of estimated marker effects for different environments at locus j, $\dot{\mathbf{Z}}_{\mathbf{j}} = \oplus_{k=1}^{K}\mathbf{z_{jk}}$, $\mathbf{z_{jk}}$ is a vector of allele dosage for locus j and environment k, and $\oplus$ denotes the direct matrix product. The update of each vector of marker effects is followed by the update of the residuals, as

$$\hat{\mathbf{e}}_{\mathbf{k}}^{(t+1)} = \hat{\mathbf{e}}_{\mathbf{k}}^{(t)} - \mathbf{Z}_{\mathbf{j}(\mathbf{k})}'(\hat{\boldsymbol{\beta}}_{\mathbf{j}(\mathbf{k})}^{(t+1)} - \hat{\boldsymbol{\beta}}_{\mathbf{j}(\mathbf{k})}^{(t)}). \tag{41}$$

Depending on the number of environments, the inversion of $\hat{\Sigma}_\beta$ may require bending (Hayes and Hill 1981; Meyer 2019), which is a technique that forces a covariance matrix to be positive-definite. Inversions can also be avoided using alternative systems of equations (see Appendix C).

Simplified covariance structures

Equation (38) provides a solution for unstructured covariance matrices. Their structure can be simplified similar to covariance matrices from extended factor analytic (XFA) models, which is meant to capture the main GxE interaction patterns (Thompson *et al.* 2003; Meyer 2009a, 2009b). The XFA covariance matrix is obtained by decomposing the unstructured $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ via EVD, and then reassembling it using a few eigenpairs that explain most of the variation while reinstating the original diagonal elements to avoid changes in heritability. Let

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbf{U}\mathbf{D}^{\mathbf{2}}\mathbf{U}', \tag{42}$$

and consider that $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{D}}$ are subsets of $\mathbf{U}$ and $\mathbf{D}$. The XFA covariance matrix is obtained with

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}(\mathbf{XFA})} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}^{\mathbf{2}}\tilde{\mathbf{U}}' + \mathbf{N}, \tag{43}$$

where $\mathbf{N}$ is a diagonal matrix that recovers the original diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.

Heteroskedastic compound symmetry (HCS) is another covariance structure attainable by the simplification of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. In an HCS structure, all pairs of environments have the same GxE correlation, and environments have different variances. Here, this structure is derived within the PEGS algorithm from $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ by calculating the average GxE correlation of pairs of environments and then covariances of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}(\mathbf{HCS})}$ are calculated using equation (5).

Benchmark analyses

Sparse testing benchmark

One hundred environments were simulated to test the predictive performance with varying amounts of missing values per environment (75% and 90%), levels of heritability ($h^2 = \{0.2, 0.4, 0.6\}$), and levels of correlation among environments ($\rho_{\mathrm{GxE}} = \{0.2, 0.4, 0.6\}$). Two genetic models for simulating phenotypes were used: (1) HCS with a constant GxE correlation and (2) an unstructured MV model with different GxE correlations between environments. For the latter, correlations ranged from −0.5 to 0.8, with a mean of 0.0 and a SD of 0.3. The phenotypes were simulated using the function `SimGC` from the R package bWGR (Xavier *et al.* 2019).

The accuracy was measured as the correlation between true and estimated breeding values within environments. The simulated dataset used real genomic information from 9 soybean families from the SoyNAM panel (Song *et al.* 2017; Diers *et al.* 2018; Xavier *et al.* 2018), where each family consists of ∼140 fullsibs genotyped with 4,240 markers.

MegaLMM is implemented in the R package MegaLMM (Runcie *et al.* 2021), and the other prediction methods are implemented in the R package bWGR (Xavier *et al.* 2019). The function `MRR3F` was used to run the UV and MV models with unstructured, HCS and XFA covariance structures. The function `ZSEMF` was used to run the full-rank MegaSEM [equation (19)]. Three eigenpairs were used to construct the XFA covariance matrices. Coefficients and variance components were estimated with PEGS for all methods except for MegaLMM, which uses Bayesian Gibbs sampling (BGS).

Runtime benchmark

Runtime was assessed by 5 simulated scenarios that vary the number of individuals from a biparental family ($n = \{500, 2{,}000, 20{,}000\}$) and the number of environments ($k = \{10, 50, 200, 2{,}000\}$). The within-environment heritability was set to 0.2. Simulated covariance matrices were unstructured, with GxE correlations ranging from −0.5 to 0.8 with a mean of

0.0 and a SD of 0.3. The genomic information was based on 10 chromosomes of 500 cM with 1 marker per cM. Accuracy was measured as the mean correlation between predicted and true breeding values within environments. Genomic information, correlation matrices, and phenotypes were simulated using the functions `SimZ`, `SimGC`, and `SimY` from the R package bWGR (Xavier *et al.* 2019), respectively. Appendix E provides an example of the simulation.

The complete dataset allows us to assess approaches that require balanced testing, including diagonalization [equation (30)] and CT [equations (6), (7), and (8)]. The following methods were evaluated: REML-based genomic best linear unbiased prediction (GREML) and its diagonalized version (D-GREML), MegaLMM, MegaSEM, UV by environment, MV models with unstructured, XFA and HCS covariance structures, and SCT. GREML, D-GREML, and MegaLMM were parameterized with genomic relationships, whereas other methods were fit as ridge regression [equation (4)], using the SVD parameterization [equation (28)] in scenarios with more markers than individuals. MegaSEM, UV, MV, XFA, HCS, and SCT utilized the PEGS solver, whereas MegaLMM was fitted using BGS. GREML and D-GREML were fit using asreml-r 4.2 (Gilmour *et al.* 2017). REML-based approaches were not tested for scenarios with more than 50 environments due to the runtime and lack of algorithmic stability. MegaLMM was not evaluated in the scenario with 20,000 individuals due to computational requirements, and XFA was not evaluated in the scenario with 2,000 traits due to numeric instability.

Real data benchmark

Predictive ability was evaluated with a corn dataset from the 2022 Genomes-to-Fields (G2F) GXE prediction competition. This dataset provides grain yield from 217 locations in the United States, harvested from 2014 to 2021, and from 4,836 unique hybrids. The validation dataset contains phenotypes from 23 locations harvested in 2022 and 548 hybrids that were not observed in the training dataset. SNP genotypes were available for hybrids. For this study, a subset of 10,000 SNPs was utilized.

The following methods were evaluated: MegaLMM, MegaSEM, univariate model fitted within each environment (UVW), univariate model that fitted BLUEs estimated across all environments (UVA), MV models with unstructured, XFA, and HCS covariance structure. These models were evaluated based on predictive ability calculated as the correlation between predicted and observed values for individuals in the validation dataset. We distinguished 3 different predictive ability values:

1) *Pairwise*: Average of by-environment correlation between predicted and observed values of an environment. This metric informs the expectation of how any given environment from the training dataset would predict any given environment from the validation dataset.
2) *Region*: Predictions were averaged across environments within a state and then correlated with validation phenotypes from environments of the same state. This metric assesses how the average of environments located in close proximity can predict a new environment within the same geographical scope.
3) *Overall*: Predictions were averaged across all environments in the training dataset and correlated with phenotypes from each environment in the validation set. This metric assesses how well averages across all environments, which is thought of as a proxy of the targeted population of environments (TPE), predict a new environment.

Results and discussion

Sparse testing benchmark

Results are presented in Fig. 1. Heritability is the most influential factor for the accuracy of predictions. It was not influenced by the proportion of missing data, based on the 2 simulation parameters (75% and 90%). The difference in the accuracy of predictions between UV and methods that capture GxE interactions increased with increasing GxE correlations. The accuracy of UV did not vary with changes in GxE correlation because the GxE information is not captured when each environment is analyzed separately.

XFA was among the top predictive models in all scenarios. The HCS model had the highest accuracy in scenarios with constant, positive GxE correlation because it was the model used to simulate phenotypes, but HCS was equivalent to UV in the unstructured GxE scenario. MegaLMM was a top-performing model in scenarios with constant, positive GxE, and slightly lower performing in the unstructured scenario with heritability of 0.2 and 0.4. MV was a top-performing model for the unstructured GxE scenario, but was the least predictive model in the scenario with a constant GxE correlation of 0.2, and had intermediate performance when the constant GxE correlation was 0.4 and 0.6. MegaSEM provided a predictive performance between UV and MV.

In the unstructured simulation scenario, when MV was the true model, XFA and MegaLMM had similar accuracies as MV. The reason may be that covariances estimated by MV had decreasing precision with decreasing heritability and true GxE correlation. Thus, lowering the dimensionality of the covariance structure reduces the number of parameters to be estimated, which may reduce the impact of this precision on the accuracy of GEBVs. This trend was also observed by Xavier and Habier (2022).

Runtime benchmark

Results are displayed in Tables 1 and 2. All methods were faster than GBLUP solved by REML. Diagonalization considerably decreased the runtime of GBLUP, but its accuracy was not included in Table 2 due to odd results caused by singularities in the Average Information matrix. MegaLMM had a longer runtime than methods solved by PEGS because it uses an MCMC solver. MV took longer than HCS and XFA in the scenario with 200 and 2,000 environments. The lowest runtimes were obtained by UV, SCT, and MegaSEM. SCT provided approximately the same runtime as UV because it consists of running UV models in transformed spaces. The runtime of MegaLMM was more sensitive to the number of individuals than the number of traits. The runtime of MV, XFA, and HCS were more sensitive to the number of environments.

The accuracy of UV was insensitive to the number of environments, as it does not capture any GxE information. All methods that capture GxE information were as predictive or better than UV, although the difference in the accuracy of GEBVs declined as the number of individuals increased or GxE correlation decreased. In scenarios with 10 environments, only SCT and MV provided the same accuracy as GBLUP but the accuracy of SCT decreased as the number of environments increased. MV was the most accurate model in all scenarios under 200 environments but its accuracy dropped in the scenario with 2,000 environments, due to the number of parameters estimated in $\mathbf{\Sigma}_{\boldsymbol{\beta}}$ and a need for bending this matrix to obtain its inverse. In the scenario with 2,000 environments, the highest accuracy was obtained by MegaLMM followed by MegaSEM. MegaSEM provided either the highest or second-highest accuracy in all scenarios except the one with the lowest dimensionality.

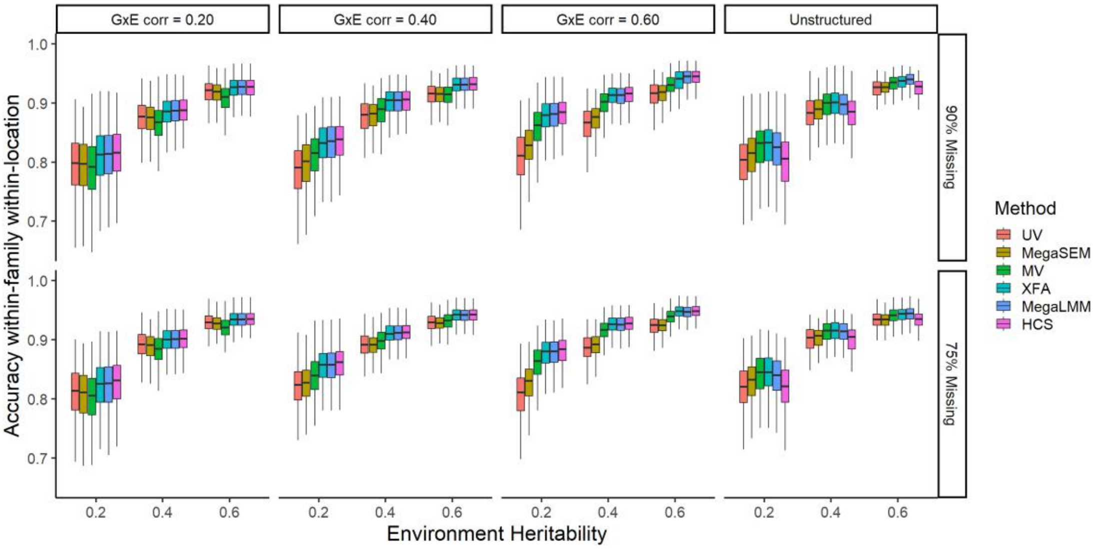


Fig. 1. Prediction accuracy within-family and within-environment using 100 simulated environments with varying heritability, percentage of missing values, and GxE correlation. Genomic information was sourced from 9 soybean biparental families.

Table 1. Average runtime in minutes (SE) for the balanced experimental design based on 10 simulated replicates.

Model	Solver	10/500	10/2,000	50/2,000	200/2,000	2,000/2,000	200/20,000
GREML	REML	46.75 (0.37)	172.61 (17.93)	—	—	—	—
D-GREML	REML	0.06 (<0.1)	0.19 (<0.1)	8.32 (3.51)	—	—	—
MegaLMM	MCMC	0.31 (0.01)	4.38 (0.06)	7.23 (1.19)	17.71 (4.02)	130.77 (11.51)	—
MegaSEM	PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.01)	2.92 (0.02)	5.26 (0.07)
MV	PEGS	<0.01 (<0.01)	<0.1 (<0.01)	0.02 (<0.01)	9.12 (1.62)	97.14 (1.29)	82.22 (5.71)
XFA	PEGS	<0.01 (<0.01)	<0.1 (<0.01)	0.03 (<0.01)	0.49 (0.09)	—	81.46 (1.38)
HCS	PEGS	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (<0.01)	0.22 (0.04)	38.74 (3.60)	37.74 (4.45)
SCT	PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.15 (0.01)	1.65 (0.01)	5.25 (0.05)
UV	PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.01)	1.44 (0.01)	5.20 (0.06)

Six scenarios vary in terms of the number of environments and individuals (no. environments/no. individuals). Models are ordered based on computational performance. The SE is shown in parenthesis.

Table 2. Within environment accuracy for the balanced experimental design based on 10 simulated replicates.

Model	Solver	10/500	10/2,000	50/2,000	200/2,000	2,000/2,000	200/20,000
GREML	REML	0.81 (0.03)	0.89 (<0.01)	—	—	—	—
MegaLMM	MCMC	0.78 (0.04)	0.87 (<0.01)	0.87 (<0.01)	0.89 (<0.01)	0.90 (<0.01)	—
MegaSEM	PEGS	0.79 (0.04)	0.88 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	0.96 (<0.01)
MV	PEGS	0.81 (0.03)	0.89 (<0.01)	0.89 (<0.01)	0.90 (<0.01)	0.88 (<0.01)	0.96 (<0.01)
XFA	PEGS	0.80 (0.04)	0.89 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	—	0.96 (<0.01)
HCS	PEGS	0.81 (0.03)	0.88 (<0.01)	0.88 (<0.01)	0.88 (<0.01)	0.88 (<0.01)	0.96 (<0.01)
SCT	PEGS	0.81 (0.03)	0.89 (<0.01)	0.88 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.95 (<0.01)
UV	PEGS	0.78 (0.04)	0.87 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.95 (<0.01)

Six scenarios vary in terms of the number of environments and individuals (no. environments/no. individuals). Models are ordered based on computational performance. SE is shown in parenthesis.

When taking into account both runtime and accuracy, our results indicate that the best method depends on the dimensionality of the data. MegaSEM suits scenarios with a large number of individuals and traits, providing high accuracy and low runtime. SCT and diagonalized GBLUP should be considered when data are balanced and the number of environments is modest. MegaLMM suits datasets with thousands or more traits but with a moderate number of individuals. HCS, MV, and XFA are suitable for datasets with up to 200 environments.

Real data benchmark

Results are displayed in Table 3. The predictive ability of overall averages was always greater than regional averages, indicating low genotype-by-region interactions. For overall averages, the highest predictive abilities were obtained by UVA, XFA, and HCS, whereas MegaLMM and MegaSEM had the same predictive ability as UVW.

Results from real data aligned with sparse testing simulations with moderate GxE (Fig. 1). This was supported by a GxE

Table 3. Predictive ability from the 2022 G2F GxE prediction competition.

Model	Pairwise	Region	Overall
UVW	0.08 (0.03)	0.22 (0.14)	0.27 (0.11)
MV	0.12 (0.05)	0.27 (0.12)	0.30 (0.11)
MegaSEM	0.13 (0.05)	0.25 (0.15)	0.27 (0.11)
MegaLMM	0.18 (0.06)	0.24 (0.19)	0.27 (0.10)
XFA	0.21 (0.07)	0.31 (0.13)	0.35 (0.12)
HCS	0.24 (0.09)	0.34 (0.11)	0.36 (0.11)
UVA	—	—	0.35 (0.12)

Corn grain yield was observed in 4,836 hybrids across 217 locations (2014–2021) predicting 548 hybrids observed across 21 environments (2022). Models are ordered based on the pairwise metric. The SE is shown in parentheses.

correlation of 0.4 estimated by HCS. Because UVA was among the most predictive models, we fitted UVW on the residuals of UVA to investigate how much GxE was left from UVA. This led to a slight improvement in the predictive ability of overall averages, from 0.35 (0.12) to 0.36 (0.12), which indicates low GxE correlations after fitting the main genetic term.

The precision of estimated GxE correlations and thereby accuracies of predictions of models with more complex covariance structures are lower for the following reasons. Firstly, a large number of environments had a small number of observations, whereas 3 environments had as few as 22 individuals. Secondly, the number of individuals that overlap across environments was limited with a median overlap between pairs of environments of 19 individuals. Thirdly, the relatedness between individuals within and across environments may be not high enough to provide more accurate genetic parameter estimates. Thus, collectively, if there was a stronger variation in GxE correlations between pairs of environments, these reasons may not allow the more complex models to reliably detect it. This may be different in commercial breeding data where the relatedness between individuals and the number of individuals across environments is expected to be higher. Despite those challenges, even the models with complex covariance structures converged. In conclusion, these results help select the right model for the data structure in a given dataset.

In this study, the G2F dataset was solely evaluated based on its predictive ability. However, models that capture complex GxE interactions, such as MV, XFA, MegaSEM, and MegaLMM, have additional benefits. These models allow for environment-specific predictions, which can be used to create selection indices aimed at improving performance under specific conditions or for broad adaptation. Beyond prediction and selection, MV models provide pairwise estimates of GxE correlations and genomic heritabilities for each environment. These correlations can reveal patterns and clusters of environments, while heritability estimates inform the location quality. The insights from GxE correlations and genomic heritability estimates are valuable for planning new trials, redesigning experiments, reallocating resources, and optimizing trial networks.

Scalability of different parameterizations

A general summary of the scalability of the different parameterizations is provided in Table 4. The table shows that no method is completely scalable in all scenarios. For example, as the number of markers increases, SVD ($\mathbf{Q}\boldsymbol{\alpha}$) and genomic relationship-based methods ($\mathbf{ZZ}'$) are preferred over SNP-based regressions ($\mathbf{Z}\boldsymbol{\beta}$ and $\mathbf{Z}'\mathbf{Z}$). That is the case when genotype-by-sequencing (GBS) data are deployed. In contrast, datasets with more genotypes than markers are common when SNP arrays are utilized in experiments across multiple breeding programs and large populations (Allen et al. 2017; Song et al. 2017), as SNP regressions can provide more efficient computation of the genomic models. When the dataset contains a large number of genotypes and markers, dimensionality reduction (e.g. $\mathbf{Q}_{n|\tilde{n}}\boldsymbol{\alpha}$) provides computational feasibility without loss in accuracy, as long as there are enough principal components to capture the genetic diversity (Pocrnic et al. 2019).

Technical guidelines for modeling large datasets with multiple traits have been provided by Misztal (2008). In that study, the author recommends starting by running UV analysis and subsequently progressing to MV models. At the time, the modeling of hundreds of traits has not been considered possible because the computational cost of REML methods increases n^2 and k^3 (Misztal 2008) with the number of genotypes (n) and traits (k), while more efficient methods, such as CT, require balanced data.

BGS is an alternative to REML (Sorensen et al. 2002), as it provides a computationally stable method to estimate variance components and regression coefficients at low memory cost. However, BGS may take a long time to run as it requires a large number of MCMC samples to provide satisfying convergence. For some Bayesian methods (Jia and Jannink 2012), variational Bayesian approaches have been proposed to avoid MCMC sampling (Hayashi and Iwata 2013).

Efficient alternatives to MCMC are available when the variance components are known *a priori* and only coefficients need to be inferred, those include Preconditioned Conjugate Gradient and GS (Legarra and Misztal 2008; Misztal and Legarra 2017). Only a few software implement GS as the main approach to estimate marker effects (Legarra et al. 2011). Here, we assessed the PEGS solver from Xavier and Habier (2022) as an approach to use GS while estimating variance components. While computationally efficient, the PEGS solver does not compute accuracies or confidence intervals because the inverse of the left-hand side of mixed-model equations or samples of effects from MCMC algorithms is not available.

Prediction of unobserved environments

In Table 3, a new environment was predicted using averaged predictions from previously observed environments. However, if the covariance between the training and prediction set were known, the prediction of new environments could be predicted as a linear combination of observed environments. Such covariances may be inferred from data when GxE interactions can be explained by a set of variables that are available for both training and validation datasets, such as management and environmental variables.

A schematic evolution of methods integrating genomics and environmental information is provided by Crossa et al. (2022), including crop-growth and reaction norm models. The covariance is inferred by environmental variables and, thus confined to their sample space. Alternatively, the associations between environmental variables and marker effects can be inferred in subsequent analyses. For instance, Della Coletta et al. (2023) GxE interaction networks are generated from the correlation of principal components of marker effects and principal components of environmental variables.

Post hoc modeling of the factors responsible for GxE interactions can be built from the output of unstructured models. Unlike crop-growth and reaction norm models, it does not assume that all interactions can be explained by the environmental variables available for modeling. The approach described here works by modeling covariances and, subsequently, generates predictions using conditional expectations.

Consider a scenario where a set of individuals A, observed in a set of environments X, is used to predict a new set of individuals B

Table 4. Scalability rating by parameterization and compatible solver.

	Parameterization	Solver	No. of genotypes	No. of markers	No. of traits
1	$\mathbf{Z}'\mathbf{Z}$	REML/BGS	****	*	*
2	$\mathbf{ZZ}'$, $\mathbf{K}$ [equation (20)]	REML/BGS	**	****	*
3	$\mathbf{Z}\boldsymbol{\beta}$ (BayesABC)	BGS	**	**	*
4	$\mathbf{U}\boldsymbol{\theta}$ [equation (30)]	REML/BGS	**	****	**
5	$\mathbf{Y}\boldsymbol{\Psi}$ [equation (13)]	BGS	**	**	**
6	$\mathbf{Q}\boldsymbol{\alpha}$ [equations (22) and (28)]	PEGS	**	***	***
7	$\mathbf{Q}_{\tilde{n}\mid n}\boldsymbol{\alpha}$ [equation (27)]	PEGS	****	***	***
8	$\mathbf{Z}\boldsymbol{\beta}$ [equations (4) and (31)]	PEGS	**	**	***
9	$\mathbf{F}\boldsymbol{\lambda}$ [equation (10)]	BGS	*	****	****
10	$\mathbf{F_0}\boldsymbol{\alpha}$ [equation (19)]	PEGS	***	***	****

observed in a set of environments Z. The estimated marker effects from prediction models are based on the observed data AX. The prediction of B individuals in observed environments is given by

$$\hat{\mathbf{G}}_{BX} = \mathbf{Z}_B\hat{\mathbf{B}}_{AX}, \tag{44}$$

where $\hat{\mathbf{G}}_{BX}$ is the matrix of GEBV of B individuals on X environments, $\mathbf{Z}_B$ is the marker information for B individuals, and $\hat{\mathbf{B}}_{AX}$ is the matrix of marker effects for X environments. The next step consists of projecting B individuals into Z environments. That is attained with the conditional expectation, where Z environments are predicted from a linear combination of X environments. Thus,

$$\hat{\mathbf{G}}_{BZ\mid BX} = \boldsymbol{\Sigma}_{ZX}\hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\mathbf{G}}_{BX}, \tag{45}$$

where $\boldsymbol{\Sigma}_X$ is the genetic variance–covariance matrix of X environments, and $\boldsymbol{\Sigma}_{ZX}$ is the covariance matrix between X and Z environments. Note that $\boldsymbol{\Sigma}_X$ is estimated from the MV model [equation (38)] that estimated the marker effects, since $\mathbf{B}_{AX} \sim N(0, \boldsymbol{\Sigma}_X \otimes \mathbf{I})$. Estimating σ_{ZX} requires prediction if Z has not been observed.

The prediction of $\boldsymbol{\Sigma}_{ZX}$ can be inferred from parameters that drive GxE interactions. Let

$$\begin{aligned}\boldsymbol{\Sigma}_X &= \mathbf{U}_X\mathbf{D}_X^2\mathbf{U}_X' \\ &= \mathbf{Q}_X\mathbf{Q}_X'\end{aligned} \tag{46}$$

where $\mathbf{Q}_X = \mathbf{U}_X\mathbf{D}_X$. Now, assuming that the principal components $\mathbf{Q}_X$ can be modeled as a linear function of parameters that drive GxE interactions ($\mathbf{W}_X$), we obtain

$$\mathbf{Q}_X = \mathbf{W}_X\boldsymbol{\Omega}_X + \mathbf{E}_X \tag{47}$$

where $\mathbf{W}_X$ is the design matrix of explanatory variables, $\boldsymbol{\Omega}_X$ is the matrix of regression coefficients, $\mathbf{E}_X$ is the matrix of residuals. Note that equation (47) acts as a post hoc modeling of environmental reaction norms, such that if the same set of variables is known for environments Z, principal components can be predicted using $\mathbf{W}_Z$. Thus,

$$\hat{\mathbf{Q}}_{ZX} = \mathbf{W}_Z\hat{\boldsymbol{\Omega}}_X \tag{48}$$

and the covariance between environments X and Z can be inferred by

$$\hat{\boldsymbol{\Sigma}}_{ZX} = \hat{\mathbf{Q}}_{ZX}\mathbf{Q}_X'. \tag{49}$$

Note that equation (47) utilizes a linear model to fit the eigenstructure of the GxE covariance; however, the evaluation of nonparametric models (e.g. random forest) is encouraged when the interaction patterns are complex beyond additivity (Alves *et al.* 2020; Waters *et al.* 2023; Resende *et al.* 2024).

Conclusion

Scalable MV approaches increase the accuracy of GEBVs within environments compared to a UV approach. Specialized models, parameterizations, and solvers enable an increasing number of individuals, markers, and environments.

In the sparse testing simulation, XFA and MegaLMM were the most accurate methods across scenarios, and HCS when GxE was constant. In the balanced simulation where runtime and accuracy were recorded, the MV was the most accurate model up to 200 environments, then surpassed by MegaLMM in the scenario with 2,000 environments. PEGS-based models were considerably more efficient than REML-based GBLUP, where MegaSEM, SCT, and UV provided the lowest runtime. In the real data analysis, predictions from UVA and overall prediction averages from HCS and XFA were the most predictive approaches. The capability of MV and MegaSEM to capture unstructured GxE patterns did not translate in higher accuracy than models with simpler covariance structures. Future studies should consider fitting multiple trait–environment combinations.

Data availability

Soybean genomic data utilized to simulate sparse testing is available in the R package `mas`, also available in the R package SoyNAM and project website: https://www.soybase.org/SoyNAM/. Corn Genomes-to-Field dataset utilized for real data benchmark is available on the project website: https://www.maizegxeprediction.org/. Data and R scripts to reproduce simulations are available on GitHub: https://github.com/alenxav/GXE24.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflicts of interest

The author(s) declare no conflicts of interest.

Literature cited

Allen AM, Winfield MO, Burridge AJ, Downie RC, Benbow HR, Barker GL, Wilkinson PA, Coghill J, Waterfall C, Davassi A, *et al.* 2017. Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). Plant Biotechnol J. 15(3):390–401. https://doi.org/10.1111/pbi.2017.15.issue-3

Alves RS, de Resende MDV, Azevedo CF, Silva FFe, Rocha JRASC, Nunes ACP, Carneiro APS, dos Santos GA. 2020. Optimization of *Eucalyptus* breeding through random regression models allowing for reaction norms in response to environmental gradients. Tree Genet Genomes. 16(1):1–8. https://doi.org/10.1007/s11295-020-01431-5

Bermann M, Lourenco D, Forneris NS, Legarra A, Misztal I. 2022. On the equivalence between marker effect models and breeding value models and direct genomic values with the algorithm for proven and young. Genet Sel Evol. 54(1):52. https://doi.org/10.1186/s12711-022-00741-7

Bustos-Korts D, Malosetti M, Chapman S, van Eeuwijk F. 2016. Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics, In: Crop Systems Biology: Narrowing the Gaps between Crop Modelling and Genetics. Springer. p. 55–82.

Crossa J, Fritsche-Neto R, Montesinos-Lopez OA, Costa-Neto G, Dreisigacker S, Montesinos-Lopez A, Bentley AR. 2021. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. Front Plant Sci. 12:651480. https://doi.org/10.3389/fpls.2021.651480

Crossa J, Montesinos-Lopez OA, Pérez-Rodríguez P, Costa-Neto G, Fritsche-Neto R, Ortiz R, Martini JW, Lillemo M, Montesinos-Lopez A, Jarquin D, *et al.* 2022. Genome and environment based prediction models and methods of complex traits incorporating genotype × environment interaction. In: Genomic Prediction of Complex Traits: Methods and Protocols. Springer. p. 245–283.

Cuevas J, Crossa J, Soberanis V, Pérez-Elizalde S, Pérez-Rodríguez P, Campos G, Montesinos-López O, Burgueño J. 2016. Genomic prediction of genotype × environment interaction kernel regression models. Plant Genome. 9(3). https://doi.org/10.3835/plantgenome2016.03.0024

Della Coletta R, Liese SE, Fernandes SB, Mikel MA, Bohn MO, Lipka AE, Hirsch CN. 2023. Linking genetic and environmental factors through marker effect networks to understand trait plasticity. Genetics. 224(4):iyad103. https://doi.org/10.1093/genetics/iyad103

de Los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb). 92(4):295–308. https://doi.org/10.1017/S0016672310000285

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 193(2):327–345. https://doi.org/10.1534/genetics.112.143313

Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V, Graef G, Nelson R, Schapaugh W, Wang D, *et al.* 2018. Genetic architecture of soybean yield and agronomic traits. G3 (Bethesda). 8:3367–3375. https://doi.org/10.1534/g3.118.200332

Elias AA, Robbins KR, Doerge R, Tuinstra MR. 2016. Half a century of studying genotype × environment interactions in plant breeding experiments. Crop Sci. 56:2090–2105. https://doi.org/10.2135/cropsci2015.01.0061

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

Falconer DS, Mackay TF. 1983. Quantitative Genetics. Longman.

Gianola D, Sorensen D. 2004. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics. 167:1407–1424. https://doi.org/10.1534/genetics.103.025734

Gilmour A, Butler D, Cullis B, Gogel B, Thompson R. 2017. Asreml-r reference manual version 4. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK.

Habier D, Fernando R, Dekkers JC. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177:2389–2397. doi:https://doi.org/10.1534/genetics.107.081190

Habier D, Fernando RL, Garrick DJ. 2013. Genomic blup decoded: a look into the black box of genomic prediction. Genetics. 194: 597–607. doi:https://doi.org/10.1534/genetics.113.152207

Hardner C. 2017. Exploring opportunities for reducing complexity of genotype-by-environment interaction models. Euphytica. 213: 248. doi:https://doi.org/10.1007/s10681-017-2023-0

Hayashi T, Iwata H. 2013. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics. 14:1–14. https://doi.org/10.1186/1471-2105-14-34

Hayes J, Hill W. 1981. Modification of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics. 37:483–493. https://doi.org/10.2307/2530561

Heslot N, Akdemir D, Sorrells ME, Jannink JL. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet. 127(2):463–480. https://doi.org/10.1007/s00122-013-2231-5

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, *et al.* 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet. 127(3): 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jia Y, Jannink JL. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 192(4): 1513–1522. https://doi.org/10.1534/genetics.112.144246

Konstantinov K, Erasmus G. 1993. Using transformation algorithms to estimate (co) variance components by REML in models with equal design matrices. S Afr J Anim Sci. 23:187–191.

Legarra A, Misztal I. 2008. Computing strategies in genome-wide selection. J Dairy Sci. 91(1):360–366. https://doi.org/10.3168/jds.2007-0403

Legarra A, Ricard A, Filangi O. 2011. Gs3: Genomic Selection, Gibbs Sampling, Gauss-Seidel (and Bayescπ). INRA.

Ma A, Needell D, Ramdas A. 2015. Convergence properties of the randomized extended Gauss–Seidel and Kaczmarz methods. SIAM J Matrix Anal Appl. 36(4):1590–1604. https://doi.org/10.1137/15M1014425

Malosetti M, Ribaut JM, van Eeuwijk FA. 2013. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Front Physiol. 4:37433. https://doi.org/10.3389/fphys.2013.00044

Martini JW, Crossa J, Toledo FH, Cuevas J. 2020. On Hadamard and Kronecker products in covariance structures for genotype × environment interaction. Plant Genome. 13:e20033. https://doi.org/10.1002/tpg2.20033

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157: 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Meyer K. 1985. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. Biometrics. 41:153–165. https://doi.org/10.2307/2530651

Meyer K. 2009a. Factor-analytic models for genotype × environment type problems and structured covariance matrices. Genet Sel Evol. 41:1–11. https://doi.org/10.1186/1297-9686-41-21

Meyer K. 2009b. Factor-analytic models for genotype × environment type problems and structured covariance matrices. Genet Sel Evol. 41(1):1–11. https://doi.org/10.1186/1297-9686-41-21

Meyer K. 2019. "Bending" and beyond: better estimates of quantitative genetic parameters? J Anim Breed Genet. 136:243–251. https://doi.org/10.1111/jbg.2019.136.issue-4

Misztal I. 2008. Reliable computing in estimation of variance components. J Anim Breed Genet. 125:363–370. https://doi.org/10.1111/jbg.2008.125.issue-6

Misztal I, Legarra A. 2017. Invited review: efficient computation strategies in genomic selection. Animal. 11:731–736. https://doi.org/10.1017/S1751731116002366

Möhring J, Piepho HP. 2009. Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49:1977–1988. https://doi.org/10.2135/cropsci2009.02.0083

Montesinos-López A, Montesinos-López OA, Montesinos-López JC, Flores-Cortes CA, de la Rosa R, Crossa J. 2021. A guide for kernel generalized regression methods for genomic-enabled prediction. Heredity (Edinb). 126:577–596. https://doi.org/10.1038/s41437-021-00412-1

Ødegård J, Indahl U, Strandén I, Meuwissen TH. 2018. Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genet Sel Evol. 50:1–12. https://doi.org/10.1186/s12711-018-0373-2

Piepho HP, Möhring J, Schulz-Streeck T, Ogutu JO. 2012. A stage-wise approach for the analysis of multi-environment trials. Biom J. 54: 844–860. https://doi.org/10.1002/bimj.v54.6

Pocrnic I, Lourenco DA, Masuda Y, Misztal I. 2016. Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. Genet Sel Evol. 48: 1–9. https://doi.org/10.1186/s12711-016-0261-6

Pocrnic I, Lourenco DA, Masuda Y, Misztal I. 2019. Accuracy of genomic blup when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. Genet Sel Evol. 51:1–10. https://doi.org/10.1186/s12711-019-0516-0

Resende RT, Xavier A, Silva PIT, Resende MP, Jarquin D, Marcatti GE. 2024. Gis-based G × E modeling of maize hybrids through enviromic markers engineering. New Phytologist. 1. https://doi.org/10.1111/nph.19951

Runcie DE, Qu J, Cheng H, Crawford L. 2021. MegaLMM: mega-scale linear mixed models for genomic predictions with thousands of traits. Genome Biol. 22(1):1–25. https://doi.org/10.1186/s13059-021-02416-w

Schaeffer L. 1986. Pseudo expectation approach to variance component estimation. J Dairy Sci. 69(11):2884–2889. https://doi.org/10.3168/jds.S0022-0302(86)80743-3

Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, Song BH, Charles An YQ, Hyten D, Nelson R, *et al.* 2017. Genetic characterization of the soybean nested association mapping population. Plant Genome. 10(2):1–14. https://doi.org/10.3835/plantgenome2016.10.0109

Sorensen D, Gianola D, Gianola D. 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. Springer.

Strandén I, Garrick D. 2009. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 92(6):2971–2975. https://doi.org/10.3168/jds.2008-1929

Thompson R, Cullis B, Smith A, Gilmour A. 2003. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust N Z J Stat. 45(4):445–459. https://doi.org/10.1111/anzs.2003.45.issue-4

Thompson E, Shaw R. 1990. Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics. 46(2):399–413. https://doi.org/10.2307/2531445

Valente BD, Rosa GJ, Gianola D, Wu XL, Weigel K. 2013. Is structural equation modeling advantageous for the genetic improvement of multiple traits? Genetics. 194:561–572. https://doi.org/10.1534/genetics.113.151209

VanRaden P, Jung Y. 1988. A general purpose approximation to restricted maximum likelihood: the tilde-hat approach. J Dairy Sci. 71:187–194. https://doi.org/10.3168/jds.S0022-0302(88)79541-7

Waters DL, van der Werf JH, Robinson H, Hickey LT, Clark SA. 2023. Partitioning the forms of genotype-by-environment interaction in the reaction norm analysis of stability. Theor Appl Genet. 136:99. https://doi.org/10.1007/s00122-023-04319-9

Xavier A, Habier D. 2022. A new approach fits multivariate genomic prediction models efficiently. Genet Sel Evol. 54:1–15. https://doi.org/10.1186/s12711-022-00730-w

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, Beavis WD, Diers BW, Song Q, Cregan PB, *et al.* 2018. Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3 (Bethesda). 8:519–529. https://doi.org/10.1534/g3.117.300300

Xavier A, Muir W, Rainey K. 2019. bWGR: Bayesian whole-genome regression. Bioinformatics. 36(6):1957–1959. https://doi.org/10.1093/bioinformatics/btz794

Xu S. 2003. Theoretical basis of the Beavis effect. Genetics. 165: 2259–2268. doi:10.1093/genetics/165.4.2259

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 44:821–824. https://doi.org/10.1038/ng.2310

Zhou X, Stephens M. 2014. Efficient multivariate linear mixed-model algorithms for genome-wide association studies. Nat Methods. 11(4):407–409. https://doi.org/10.1038/nmeth.2848

Appendix A: Canonical transformation

Under CT, the matrix of phenotypes is converted into a series of orthogonal canonical traits (**F**) as

$$\mathbf{F} = \mathbf{Y}\mathbf{M} \tag{A1}$$

where **M** has the following properties:

$$\begin{aligned} \mathbf{M}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{M}' &= \mathbf{D}, \\ \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{M}' &= \mathbf{I}, \end{aligned} \tag{A2}$$

The matrix **M** is obtained using EVD by first decomposing $\boldsymbol{\Sigma}_{\mathbf{e}}^{\mathbf{0}} = \mathbf{U}_{\mathbf{e}}\mathbf{D}_{\mathbf{e}}^{\mathbf{2}}\mathbf{U}_{\mathbf{e}}'$, then building $\mathbf{T} = \mathbf{R}_{\mathbf{e}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\mathbf{0}}\mathbf{R}_{\mathbf{e}}'$ where $\mathbf{R}_{\mathbf{e}} = \mathbf{U}_{\mathbf{e}}\mathbf{D}_{\mathbf{e}}^{-\mathbf{1}}$, subsequently decomposing $\mathbf{T} = \mathbf{U}_{\mathbf{T}}\mathbf{D}_{\mathbf{T}}^{\mathbf{2}}\mathbf{U}_{\mathbf{T}}'$, resulting on

$$\mathbf{M} = \left(\mathbf{U}_{\mathbf{T}}\mathbf{R}_{\mathbf{e}}'\right)^{-\mathbf{1}}. \tag{A3}$$

The covariances are converted back to the original scale with

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \mathbf{M}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\mathbf{F}}\mathbf{M}', \\ \boldsymbol{\Sigma}_{\mathbf{e}} &= \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{e}}^{\mathbf{F}}\mathbf{M}'. \end{aligned} \tag{A4}$$

where covariances in transformed scale are notates as $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\mathbf{F}}$, $\boldsymbol{\Sigma}_{\mathbf{e}}^{\mathbf{F}}$ and the starting values of each iteration as $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\mathbf{0}}$, $\boldsymbol{\Sigma}_{\mathbf{e}}^{\mathbf{0}}$. CT iterates on (1) transforming the phenotypes; (2) solving the canonical trait model; and (3) transforming covariances back to the phenotypic. These steps (1–3) are repeated up to convergence ($\boldsymbol{\Sigma} \approx \Sigma^0$).

Appendix B: Connection between MegaLMM and SCT

A relation between the latent spaces from MegaLMM with the SCT can be drawn by defining the variance not capture latent spaces $\mathbf{J}$ in terms of leftover rotation, as $\mathbf{J} = \mathbf{FH}$ and $\mathbf{V}' = \boldsymbol{\Lambda} + \mathbf{H}$, such that

$$\begin{aligned}\mathbf{Y} &= \mathbf{F}\boldsymbol{\Lambda} + \mathbf{J}\\ &= \mathbf{F}\boldsymbol{\Lambda} + \mathbf{FH}\\ &= \mathbf{F}(\boldsymbol{\Lambda} + \mathbf{H})\\ &= \mathbf{FV}',\end{aligned} \tag{B1}$$

making the decomposition from equation (9) feasible and parsimonious approximation of $\mathbf{FV}'$ when it is not possible to compute the exact rotation via SVD due to the presence of missing value in the phenotypic matrix.

Appendix C: Avoiding inversions with PEGS

The nonvector operations involved in the multivariate PEGS are (1) the inversion of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ and (2) solving the marker effects [equation (40)]. The former issue can be mitigated by multiplying both sides of the equation by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$, as proposed by Strandén and Garrick (2009). Thus,

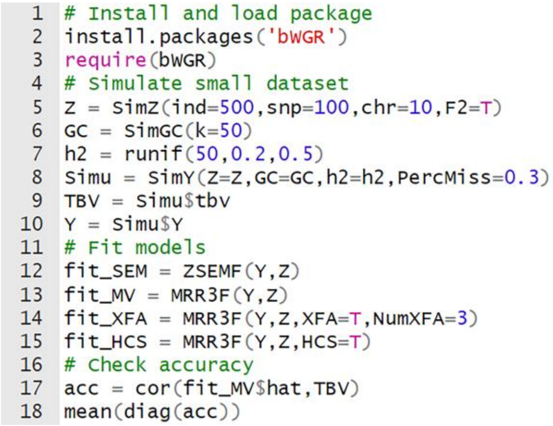
$$\begin{aligned}(\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'\dot{\mathbf{Z}}_{\mathbf{j}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{\mathbf{-1}})\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t+1})} &= \hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'(\dot{\mathbf{Z}}_{\mathbf{j}}\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t})} + \hat{\mathbf{e}})\\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'\dot{\mathbf{Z}}_{\mathbf{j}} + \hat{\boldsymbol{\Sigma}}_{\beta}^{-1})\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t+1})} &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'(\dot{\mathbf{Z}}_{\mathbf{j}}\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t})} + \hat{\mathbf{e}})\\ (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'\dot{\mathbf{Z}}_{\mathbf{j}} + \mathbf{I})\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t+1})} &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}'(\dot{\mathbf{Z}}_{\mathbf{j}}\hat{\boldsymbol{\beta}}_{\mathbf{j}}^{(\mathbf{t})} + \hat{\mathbf{e}}).\end{aligned} \tag{C1}$$

For the latter issue, one may address the inversion of the left-hand side by running an inner GS, for the single site update of coefficients. Since the algorithm convergence is computed across coefficients for all marker–environment combinations, a single iteration of inner GS per marker update suffices. Let

$$\begin{aligned}\mathbf{C} &= (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_{\mathbf{j}}{}'\dot{\mathbf{Z}}_{\mathbf{j}} + \mathbf{I}),\\ \boldsymbol{\beta} &= \hat{\boldsymbol{\beta}}_{\mathbf{j}},\\ \mathbf{r} &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\dot{\mathbf{Z}}_{\mathbf{j}}'\hat{\boldsymbol{\Sigma}}_e^{-1}(\dot{\mathbf{Z}}_{\mathbf{j}}\hat{\boldsymbol{\beta}}_{\mathbf{j}} + \hat{\mathbf{e}}),\end{aligned} \tag{C2}$$

aiming to solve $\mathbf{C}\boldsymbol{\beta} = \mathbf{r}$. The single site update of the jth marker effect, at kth trait consists of

$$\boldsymbol{\beta}_{\mathbf{k}} = \frac{\mathbf{r}_{\mathbf{k}} - \mathbf{C}_{\mathbf{k},-\mathbf{k}}\boldsymbol{\beta}_{-\mathbf{k}}}{\mathbf{C}_{\mathbf{k},\mathbf{k}}}. \tag{C3}$$



```r
# Install and load package
install.packages('bWGR')
require(bWGR)
# Simulate small dataset
Z = SimZ(ind=500,snp=100,chr=10,F2=T)
GC = SimGC(k=50)
h2 = runif(50,0.2,0.5)
Simu = SimY(Z=Z,GC=GC,h2=h2,PercMiss=0.3)
TBV = Simu$tbv
Y = Simu$Y
# Fit models
fit_SEM = ZSEMF(Y,Z)
fit_MV = MRR3F(Y,Z)
fit_XFA = MRR3F(Y,Z,XFA=T,NumXFA=3)
fit_HCS = MRR3F(Y,Z,HCS=T)
# Check accuracy
acc = cor(fit_MV$hat,TBV)
mean(diag(acc))
```

Fig. E1. Code to simulate data and fit efficient multivariate models in R using the bWGR package (version 2.2.9).

Appendix D: Sparse inversion of kernels

Kernel-based models [equation (20)] are commonly solved through the inversion of $\mathbf{K}$. When the number of genotyped individuals is large, the algorithm for proven and young (APY) provides a sparse representation of the inverse genomic relationship matrices (Pocrnic *et al.* 2016; Bermann *et al.* 2022). Under APY, dense inversion is performed only on a subset of the relationship matrix containing a representative sample of individuals referred to as the core set. Thus,

$$\mathbf{K}_{\mathbf{APY}}^{-1} = \begin{bmatrix}\mathbf{K}_{\mathbf{cc}}^{-1} + \mathbf{P}_{\mathbf{cn}}\mathbf{M}_{\mathbf{nn}}^{-1}\mathbf{P}_{\mathbf{nc}} & -\mathbf{P}_{\mathbf{cn}}\mathbf{M}_{\mathbf{nn}}^{-1}\\ -\mathbf{M}_{\mathbf{nn}}^{-1}\mathbf{P}_{\mathbf{nc}} & \mathbf{M}_{\mathbf{nn}}^{-1}\end{bmatrix}, \tag{D1}$$

where $\mathbf{c}$ and $\mathbf{n}$ describe the core and noncore set of genotypes, respectively, $\mathbf{P} = \mathbf{K}_{\mathbf{nc}}\mathbf{K}_{\mathbf{cc}}^{-1}$, and $\mathbf{M}$ is a diagonal matrix with the element-wise Schur complement, as $\mathbf{m}_{ii} = \{\mathbf{k}_{\mathbf{ii}} - \mathbf{k}_{ic}\mathbf{K}_{\mathbf{cc}}^{-1}\mathbf{k}_{ci}\}$.

Appendix E: Sample code

An example of how data was simulated and how PEGS-based models were fitted in R is provided in Fig. E1. It simulated 500 F2 individuals with 10 chromosomes of 100 SNPs. It fits MegaSEM, MV, XFA, and HCS. Last, it computed the average accuracy within the environment for one of the models as the correlations between estimated and true breeding values.

Editor: M. Calus